The Dissertation Committee for Teresa Nicole Portone certifies that this is the approved version of the following dissertation:

Representing Model-Form Uncertainty from Missing Microstructural Information

Committee:

Robert D. Moser, Supervisor

Clint Dawson

Omar Ghattas

Peter Müller

Damon McDougall

Todd A. Oliver

Representing Model-Form Uncertainty from Missing Microstructural Information

by

Teresa Nicole Portone

Dissertation

Presented to the Faculty of the Graduate School
of the University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

The University of Texas at Austin
December 2019

Dedicated to my parents, Maria and Frank Portone.

Acknowledgments

First, I would like to thank my advisor, Dr. Robert Moser, for patiently molding me into a computational scientist. I would like to thank my committee members for their support while completing this thesis, and in particular I would like to thank Drs. Damon McDougall and Todd Oliver for taking the time to answer countless questions while I was on the steepest part of my learning curve. Thank you to Prakash Mohan for always being happy to discuss research, and for sending me photos of cats to cheer me up while I was writing this dissertation. Thank you to my friends for reminding me to have a life outside of work—Sam, Gopal, Sid, Tim, Travis, Sameer, Keith, and Aaron especially. To my sister, Carissa, for your many pep talks and for your invaluable career advice. To my parents, who always believed in me, even when I did not believe in myself.

Finally, thank you to Bryan Reuter. You have been by my side through this entire experience, from getting me up to speed in our first-year Mathematical Modeling class, to reading several drafts of this dissertation. You have supported and encouraged me, and you have kept me laughing through the rough spots. I may have gotten to this point without you, but it would have been a lot less fun.

This work was supported by two U.S. Department of Energy Mathematical Multifaceted Integrated Capabilities Centers: AEOLUS and DiaMonD: An Integrated Multifaceted Approach to Mathematics at the Interfaces of Data, Models, and Decisions. Computational resources were provided by the Texas Advanced Computing Center (TACC).

Representing Model-Form Uncertainty from Missing Microstructural Information

by

Teresa Nicole Portone, Ph.D.

The University of Texas at Austin, 2019

SUPERVISOR: Robert D. Moser

A common challenge in modeling multiscale phenomena lies in representing the dependence of macroscopic quantities on microscale dynamics. Incomplete information or limitations in computational resources make it impossible to resolve the microscale dynamics and their effect on those at the macroscale. To obtain a model of the phenomenon that can be used to make predictions, approximations must be made. For instance, it is commonly assumed that microscale effects on the macroscale can be represented with macroscopic quantities, effectively removing any dependence on the microstate. Such approximations introduce uncertainty in the model. When the approximations are invalid, the uncertainty is significant and must be quantified to assess the reliability of the model. This work focuses on the formulation of a model-form uncertainty representation to account for such missing dependencies. The process by which a model-form uncertainty representation is formulated is an open area of research, so particular attention is paid to determining the feasibility and inherent challenges of its development.

The representation is developed in the context of a simplified testbed problem, accounting for uncertainty in a model of mean contaminant transport through a heterogeneous porous medium. In heterogeneous media, the evolution of the mean depends on small-scale fluc-

tuations of the flow velocity from its mean and their induced fluctuations on the detailed concentration field. However, these fluctuations can neither be observed nor resolved. In this work, model-form uncertainty caused by the unresolved dependence on the small-scale fluctuations is represented as an infinite-dimensional stochastic operator acting on the mean concentration. Physical constraints are enforced through its eigendecomposition, and uncertainty is encoded in its eigenvalues by casting them as random variables. The feasibility of inferring their mean values using observations of the mean concentration is explored, and a novel method of extracting samples from their distribution using direct numerical simulation is discussed. These findings are used to develop a stochastic model for the probability distribution of the operator's eigenvalues, and its validity is assessed using forward propagation of uncertainty to the mean concentration.

Contents

Chapter 1

Introduction

In developing models for complex problems in science and engineering it is often necessary to make simplifying assumptions and approximations. This may be for practical reasons, because available high-fidelity models are computationally intractable, or it may be due to incomplete information about the phenomenon being modeled. Approximations include, but are not limited to, assumptions of isotropy, lack of coupling between phenomena, and lack of dependence on independent variables such as time or temperature. These approximations induce uncertainty in the model's form, and when they are inaccurate the uncertainty is significant. If left unaddressed, they result in the model's inability to reproduce important observable quantities. When this occurs, the model is considered inadequate. Computational models are increasingly used to make predictions affecting high-consequence engineering-design and policy decisions. To understand the reliability of these predictions, it is essential to account for uncertainties in model form.

This is a common issue in the representation of multiscale phenomena, whose dynamics are coupled across scales. Often, macroscopic quantities of interest depend on dynamics at smaller scales. However, limitations in sensing technology and computational power make it impossible to resolve all the relevant scales in computational models of the phenomena. This inability to resolve the small-scale dynamics induces uncertainty in the models. One example for which this occurs is groundwater transport of a contaminant. In transport

through heterogeneous porous media, variations in permeability at the mesoscale (larger than the pore scale but smaller than the field scale) produce significant fluctuations in the concentration field of the contaminant. These fluctuations affect the macroscopic (field-scale) dynamics of the transport, leading to so-called "anomalous diffusion," wherein the diffusion of the contaminant profile is not well described by standard, second-order "Fickian" diffusion [1–6]. However, noninvasive observational techniques do not yet exist to observe the permeability variations with sufficient resolution over the span of an aquifer. Even if such techniques existed, it would be computationally infeasible to solve the fully-resolved equations for many practical problems. To overcome these obstacles, models of the transport at the macroscale are often homogenized, modeling away any dependence on the unobservable dynamics at the smaller scales. However, this leads to predictions of the transport that are inconsistent with observed anomalous diffusion.

In this work a physics-based uncertainty representation accounting for a macroscopic model's missing dependence on dynamics at the smaller scales is formulated. The development of model-form uncertainty representations is an active area of research, so this work focuses on assessing the feasibility and inherent challenges in its formulation. The representation is developed using a testbed problem in contaminant transport through a heterogeneous medium.

For the testbed problem, a hierarchy of model fidelities is employed. A high-fidelity model that resolves the small-scale dynamics serves as a baseline, and a model uncertainty representation is developed for a low-fidelity, homogenized version of the high-fidelity model. To enable uncertainty analysis both models are much simpler than a fully-resolved, 3D, field-scale model of contaminant transport, but the fundamental challenge of missing dependence across scales in the homogenized model persists. Because all aspects of both models are known, the only source of uncertainty is this missing dependence. Access to a high-fidelity model that resolves the small-scale dynamics allows for the phenomenon and discrepancies

between the models to be probed in a level of detail that would not be possible by comparing to experimentally-collected data. In Section 1.1 the high-fidelity model is introduced and the low-fidelity model is derived. In Section 1.2, existing approaches to addressing model-form uncertainty are discussed, and the approach taken here is described.

1.1 Testbed problem description

For the purposes of this work, the 2D advection-diffusion equation will be considered an accurate representation of a contaminant's transport through a heterogeneous porous medium. The velocity is determined by Darcy's law and the continuity equation. The governing equations are thus

$$\frac{\partial c(\mathbf{x},t)}{\partial t} + \nabla \cdot (\mathbf{u}(\mathbf{x})c(\mathbf{x},t)) = \nu_p \nabla^2 c(\mathbf{x},t), \tag{1.1}$$

$$\nabla \cdot (\mathbf{u}) = 0, \tag{1.2}$$

$$\mathbf{u}(\mathbf{x}) = -\kappa(\mathbf{x})\nabla p(\mathbf{x}), \tag{1.3}$$

where c is the concentration field of the contaminant; $\mathbf{u}$ is the velocity; ν_p represents pore-scale diffusivity; $\kappa(\mathbf{x})$ represents permeability, a measure of how easily fluid travels through the medium; and p is the pressure field. In the derivation of the homogenized model, all model parameters, initial and boundary conditions are assumed known, with the exception of κ. Because the velocity depends on κ through Darcy's law, the permeability indirectly determines the transport of the contaminant. If κ were known throughout the entire computational domain, the transport of the contaminant would be completely predictable.

In realistic problems, the structure of a permeability field for a field-scale computational domain is not available, due to limitations in sensing technology. Instead, it is possible to collect samples of the porous medium and study small sections of the domain in a laboratory. Viewing the permeability as a random field, the samples can be used to determine a mean and

correlation structure. Assuming κ is statistically homogeneous—that is, that its statistics do not depend on absolute location—the statistics determined in the lab are representative of its statistics over the whole domain. Thus, although the detailed behavior of the permeability field is not known, its statistics, and those of the velocity and other quantities that depend on it, can still be predictable.

It is common practice to make such assumptions and perform statistical averaging to derive an equation for the transport of the mean contaminant concentration field [7], and this approach is taken here. For this testbed problem κ is defined to be statistically homogeneous, so the assumption of statistical homogeneity is valid. Additionally, depthwise (y-direction) averaging is performed for two reasons. First, although the evolution of the contaminant is assumed to occur in 2D, observations of the contaminant are limited to a depthwise average due to mixing that occurs when drawing fluid from a well for measurement. Second, the depthwise variation of the contaminant's concentration is not generally the relevant quantity of interest; of more concern is when the average depthwise concentration of the contaminant exceeds a safe threshold downstream of some contaminant source.

To obtain a set of equations for the statistically- and spatially-averaged concentration, let

$$\langle f(x,y)\rangle \equiv \frac{1}{L_y}\int_0^{L_y} \mathbb{E}_\kappa\left[f(x,y)\right]\mathrm{d}y$$

for a random field f, where $\mathbb{E}_\kappa$ signifies an expectation over the probability space of κ. The random field f can thus be written as the sum of its mean and its deviation from that mean:

$$f = \langle f\rangle + f'.$$

Substituting this decomposition of c and $\mathbf{u} = [u, v]$ into the high-fidelity equations (1.1) and

(1.2) and applying the averaging operator to the equations gives

$$\begin{aligned} \frac{\partial \langle c \rangle}{\partial t} + \langle u \rangle \frac{\partial \langle c \rangle}{\partial x} + \frac{\partial \langle u'c' \rangle}{\partial x} &= \nu_p \frac{\partial^2 \langle c \rangle}{\partial x^2}, \\ \langle c \rangle (0,t) &= \langle c \rangle (L_x, t), \\ \langle c \rangle (x,0) &= c_0(x). \end{aligned} \tag{1.4}$$

Note that $\langle c \rangle$ is assumed periodic with period L_x, which is based on the fact that the relevant statistics affecting its evolution, namely the fluctuations in the velocity, are small compared to L_x. Furthermore, $\langle u \rangle$ is constant, since by continuity

$$0 = \frac{\partial \langle u \rangle}{\partial x} + \frac{\partial \langle v \rangle}{\partial y} = \frac{\partial \langle u \rangle}{\partial x}.$$

This system of equations is exact but unclosed because of the second-order fluctuating term $\langle u'c' \rangle$. This term is often called the dispersive flux, and $\partial \langle u'c' \rangle / \partial x$ is herein called the dispersion. A typical closure model for $\langle u'c' \rangle$ is gradient diffusion [7]

$$\langle u'c' \rangle \approx -\nu_m \frac{\partial \langle c \rangle}{\partial x},$$

where ν_m is a model diffusion coefficient. Substituting the gradient-diffusion model into (1.4) yields the advection-diffusion equation for the mean concentration,

$$\frac{\partial \langle c \rangle}{\partial t} + \langle u \rangle \frac{\partial \langle c \rangle}{\partial x} = \nu \frac{\partial^2 \langle c \rangle}{\partial x^2}, \tag{1.5}$$

where $\nu = \nu_p + \nu_m$.

Dispersion is caused by local fluctuations of the fluid velocity that transport the contaminant downstream at different speeds. The effect of this heterogeneous advection on the averaged concentration field is generally diffusive, which is why the gradient-diffusion model

is often employed. For field-scale transport through a heterogeneous porous medium, velocity fluctuations from the mean can be quite large, and the evolution of the mean concentration will depend significantly on the fluctuations, resulting in a mean evolution that is not well described by gradient diffusion, as shown in Figure 1.1. The concentration profile has a broad leading edge, meaning a higher concentration downstream than is predicted by gradient diffusion. This shows how the gradient-diffusion model can dangerously mispredict the time at which a contaminant's concentration will exceed a safe threshold downstream. Indeed, it is well known that gradient diffusion is an inadequate model for field-scale transport through heterogeneous porous media [1, 8, 9].

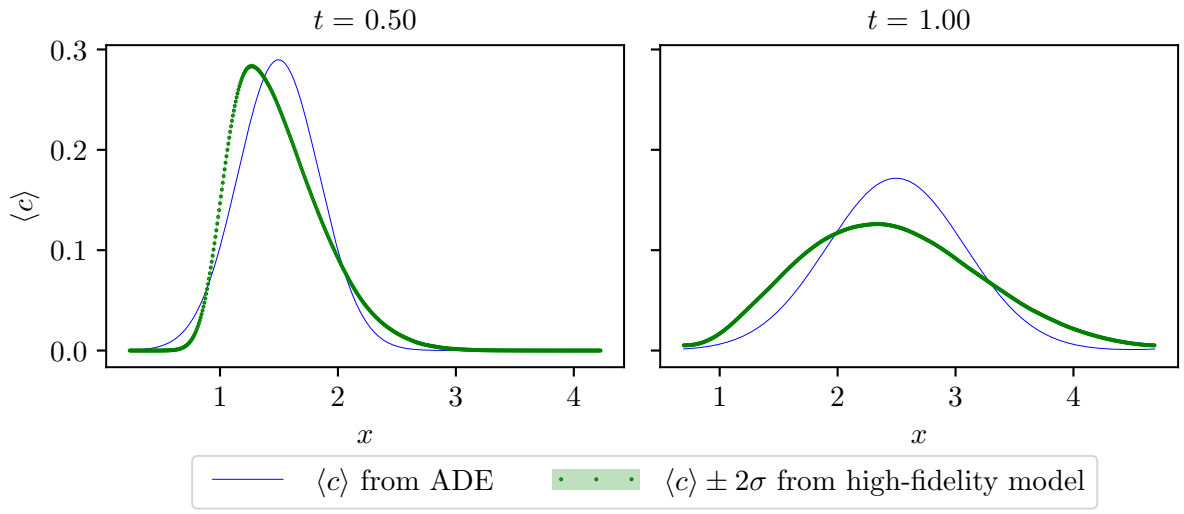

Figure 1.1: Sample mean $\langle c \rangle$ from the high-fidelity model with a 95% confidence interval from its sampling distribution, compared to the evolution of the same initial condition by advection-diffusion.

1.2 Approaches to Model Inadequacy

A model is considered inadequate if it cannot reproduce the physical phenomenon it is supposed to represent accurately enough to agree with important observable quantities in the system. In this case the gradient-diffusion closure is not able to reproduce the anomalous

diffusion of the state. Using an inadequate model induces uncertainty in its predictions. If this uncertainty is ignored, it can result in overconfidence in incorrect predictions, as shown in [10]. For this reason it is essential to address model-form uncertainty.

The concept of model inadequacy and its statistical characterization was first introduced by Kennedy and O'Hagan in [11]. In the Kennedy and O'Hagan approach to model-form uncertainty, a stochastic term (usually a Gaussian process) is appended to a data model e.g.

$$d_i = \langle c \rangle (x_i, t_i) + \epsilon_{obs} + \epsilon_m(x_i, t_i), \quad \epsilon_{obs} \sim \mathcal{N}(0, \sigma^2), \quad \epsilon_m \sim \mathcal{N}(\mu(x_i, t_i), \mathcal{C}),$$

and its mean μ and covariance $\mathcal{C}$ are calibrated to minimize misfit between model output and data. A clear downside to this approach is that it is not predictive. The model discrepancy term ϵ_m is only tuned for the specific scenarios for which there is data. Outside of this training region there are no guarantees of its effectiveness. Furthermore, there is no direct way to transfer information about the uncertainty in the observable model outputs to other quantities of interest (QOIs) the model may be used to predict.

The model discrepancy term introduced by Kennedy and O'Hagan was not intended to be used in a predictive setting—it was included to describe the inadequacy of a computational model calibrated for a specific scenario. In [10], it is assumed that the model will be used for prediction in scenarios for which calibration and validation data do not exist. This often occurs in engineering design and in cases where computational models are used to make predictions about the future. To have confidence in its predictions in such cases, the model and its uncertainty representation must be able to accurately extrapolate to prediction scenarios. Because of this, in [10] Oliver *et al.* argue that model uncertainty representations must be formulated to depend on the state variables of the model as well as key parameters that define the scenario in which the model is employed. Furthermore, they advocate embedding the uncertainty representation within the model where the important

dependencies are neglected. This embedding allows the uncertainty to be propagated to observables, so that it can be calibrated with data, as well as to important quantities of interest. Finally, they argue that the uncertainty representation should be stochastic, to represent the fact that the missing dependencies in the model cannot be resolved with any amount of available data.

In [12], Sargsyan *et al.* introduced a restricted version of the representation discussed in [10] by casting the parameters of the inadequate closure model as stochastic random variables. For the contaminant transport problem this would correspond to casting the model diffusion coefficient ν_m of the gradient-diffusion closure as a stochastic random variable. The benefit of this approach is that no additional modeling is required, beyond specifying the stochastic representation of the parameters. The downside of this approach is that the representation is limited to the dynamics achievable by the inadequate closure model. For instance, in the testbed problem, inducing stochasticity in ν_m would produce greater uncertainty in $\langle c \rangle$, but the predicted evolution would still be Fickian. With this approach, the homogenized model and its uncertainty representation still would not predict anomalous diffusion for $\langle c \rangle$.

Instead, in the approach taken in [10, 13], a physics-based stochastic representation of model uncertainty is introduced either as an augmentation or a replacement of the closure model. By defining the representation, herein denoted $\mathcal{L}$, to act on the state variable(s), it modifies the dynamics of the model in which it appears. To maximize its predictive ability, as many physical and mathematical constraints as possible are incorporated into its deterministic formulation. Remaining uncertainties are represented using probability distributions. These distributions can be updated using data, often using Bayesian inference. It is a physics-based uncertainty representation, built up from knowledge of the physical problem and the errors incurred by the assumptions made in deriving the inadequate model.

There are many ways to induce state dependence in $\mathcal{L}$. Previous approaches have included

modeling $\mathcal{L}$ as a solution to a stochastic PDE coupled with the original evolution equations and as a finite-dimensional operator acting on a vector of state variables [13]. Based on the success and intuitiveness of the finite-dimensional operator approach in [13], and because it is an approach that has not yet been studied, this work explores the feasibility of representing the model-form uncertainty as an infinite-dimensional stochastic operator acting on the mean concentration of the contaminant, $\langle c \rangle$. Another reason this approach was chosen is because it was noted that $\mathcal{L}$ must be linear to avoid violating the linearity of the conservation of mass equation with respect to $\langle c \rangle$. This meant that it would admit a tractable means of imposing constraints on its structure through its eigendecomposition.

In this work $\mathcal{L}$ will replace the unclosed dispersion term in the model of the evolution of $\langle c \rangle$ rather than augment the gradient-diffusion closure model. That is,

$$\mathcal{L}(\langle c \rangle) \approx -\frac{\partial \langle u'c' \rangle}{\partial x}.$$

Substituting this expression into (1.4) gives

$$\frac{\partial \langle c \rangle}{\partial t} + \langle u \rangle \frac{\partial \langle c \rangle}{\partial x} = \nu_p \frac{\partial^2 \langle c \rangle}{\partial x^2} + \mathcal{L}(\langle c \rangle). \tag{1.6}$$

The deterministic formulation of the stochastic operator to respect physical constraints is discussed in Chapter 2. The feasibility and challenges of inferring its mean using Bayesian inference and observations of $\langle c \rangle$ are explored in Chapter 3. A method of directly computing sample eigenvalues of the stochastic operator is described Chapter 4. The method is used to study the dependence of its distribution on scenario, defined in terms of the statistics of the permeability field. In Chapter 5, the findings from Chapter 4 are used to formulate and calibrate the stochastic representation of the operator, which is used to predict the evolution of $\langle c \rangle$. The feasibility of developing a relatively simple, inexpensive model-form uncertainty representation that accounts for a macroscopic model's missing dependence on dynamics at

smaller scales is a main focus of the work. This question and the benefits and challenges of representing the uncertainty as an infinite-dimensional stochastic operator are discussed in Chapter 6.

Chapter 2

Deterministic formulation of the inadequacy model

The first step in the development of the model uncertainty representation $\mathcal{L}$ is to impose deterministic constraints on its form based on physical and mathematical characteristics of the problem that it should respect. For instance, the mean advection-diffusion equation is linear in $\langle c \rangle$. It is also shift-invariant because of the statistical homogeneity of the underlying medium. Finally, it is an expression of conservation of mass. The deterministic formulation of $\mathcal{L}$ will be defined to respect these constraints.

First, to respect linearity in $\langle c \rangle$, $\mathcal{L}$ is defined to be a linear operator. Substituting $\mathcal{L}$ into (1.6) yields the system

$$\begin{aligned} \frac{\partial \langle c \rangle}{\partial t} + \langle u \rangle \frac{\partial \langle c \rangle}{\partial x} = \nu_p \frac{\partial^2 \langle c \rangle}{\partial x^2} + \mathcal{L} \langle c \rangle, \quad x \in (0, L_x), \\ \langle c \rangle (0,t) = \langle c \rangle (L_x, t), \\ \langle c \rangle (x,0) = c_0(x). \end{aligned} \tag{2.1}$$

Because $\mathcal{L}$ is linear, it can be specified in terms of its eigenvalues and eigenfunctions, λ_k and f_k, $k \in \mathbb{Z}$. Assuming its eigenfunctions form a basis for the solution space of (2.1), its action on $\langle c \rangle$ can thus be expressed as

$$\mathcal{L} \langle c \rangle = \sum_{k=-\infty}^{\infty} \lambda_k \langle c_k \rangle f_k, \tag{2.2}$$

where $\langle c_k \rangle$ are the expansion coefficients of $\langle c \rangle$. This parametrization enables further constraints to be applied to $\mathcal{L}$ by constraining λ_k and f_k.

The second constraint is shift invariance. Mathematically, the property of shift invariance implies that $\mathcal{L}$ should commute with the spatial shift operator $\mathcal{S}_{x'} f(x) = f(x + x')$. The solution space of (2.1) is the set of continuously-differentiable periodic functions on the bounded domain $[0, L_x]$. On this domain the shift operator's eigenfunctions are the Fourier modes, $\exp(ia_k x)$, where $a_k = 2\pi k / L_x$, $k \in \mathbb{Z}$. This implies that the Fourier modes are the eigenfunctions of $\mathcal{L}$ as well, since operators that commute share eigenfunctions. Let the Fourier coefficients of $\langle c \rangle$ be denoted $\langle \hat{c}_k \rangle$. Then the action of $\mathcal{L}$ on $\langle c \rangle$ can be expressed in a Fourier series as

$$\mathcal{L} \langle c \rangle = \sum_{k=-\infty}^{\infty} \lambda_k \langle \hat{c}_k \rangle \exp(ia_k x).$$

Since the eigenfunctions of $\mathcal{L}$ are known, only its eigenvalues λ_k are uncertain and are constrained further. The advection-diffusion equation is a statement of mass conservation, so λ_k must be defined so that $\mathcal{L}$'s action does not add or remove mass from the system. Then

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}t} \int_0^{L_x} \langle c \rangle \, \mathrm{d}x = \int_0^{L_x} \frac{\partial \langle c \rangle}{\partial t} \mathrm{d}x \\
&= \int_0^{L_x} \nu_p \frac{\partial^2 \langle c \rangle}{\partial x^2} + \mathcal{L} \langle c \rangle - \langle u \rangle \frac{\partial \langle c \rangle}{\partial x} \mathrm{d}x \\
&= \sum_{k=-\infty}^{\infty} \int_0^{L_x} \left(-\nu_p a_k^2 + \lambda_k - \langle u \rangle \, i a_k \right) \langle \hat{c}_k \rangle \, e^{i a_k x} \mathrm{d}x \\
&= \lambda_0 \langle \hat{c}_0 \rangle .
\end{aligned}$$

Thus it is sufficient to require $\lambda_0 = 0$ to preserve mass.

Finally, the mean concentration is known to decay with time as the contaminant is diffused throughout the domain. To ensure the solution decays with time it is sufficient

to guarantee that $|\langle\hat{c}_k\rangle|(t) \leq |\langle\hat{c}_k\rangle|(0)\ \forall k$, where $\langle\hat{c}_k\rangle(0)$ are the Fourier coefficients of the initial condition. The Fourier coefficients of the solution $\langle c\rangle$ to (2.1) are defined as

$$\langle\hat{c}_k\rangle(t) = \langle\hat{c}_k\rangle(0)\exp\left(\left(-\nu_p a_k^2 + \lambda_k - \langle u\rangle i a_k\right)t\right), \quad k \in \mathbb{Z}.$$

Separating this into its real and imaginary parts yields

$$\langle\hat{c}_k\rangle(t) = \langle\hat{c}_k\rangle(0)\exp\left(\left(-\nu_p a_k^2 + \Re[\lambda_k]\right)t + i\left(\Im[\lambda_k] - \langle u\rangle a_k\right)t\right), \quad k \in \mathbb{Z}.$$

Only the real part of the argument in the exponential affects the coefficients' magnitude, so

$$|\langle\hat{c}_k\rangle(t)| = |\langle\hat{c}_k\rangle(0)|\exp\left(\left(-\nu_p a_k^2 + \Re[\lambda_k]\right)t\right).$$

Then it is sufficient to require that $-\nu_p a_k^2 + \Re[\lambda_k] \leq 0$ to guarantee the solution decays with time.

To this point the physical constraints placed on the operator resulted in simple constraints on its structure that were easy to impose. A further constraint on $\mathcal{L}$ that it preserve the positivity of the concentration field was pursued, but determining a constructive way of enforcing this property proved challenging. A common approach to enforcing positivity is to instead model the natural logarithm of the positive quantity. However, the governing equation would lose linearity with respect to the state, which is a desirable property of the testbed problem. This can be seen by defining $\langle c\rangle = e^f$ and substituting it into the

advection-diffusion equation:

$$\begin{aligned} 0 &= \frac{\partial e^f}{\partial t} + \langle u \rangle \frac{\partial e^f}{\partial x} - \nu_p \frac{\partial^2 e^f}{\partial x^2} \\ &= e^f \frac{\partial f}{\partial t} + \langle u \rangle e^f \frac{\partial f}{\partial x} - \nu_p \left[e^f \left(\frac{\partial f}{\partial x} \right)^2 + e^f \frac{\partial^2 f}{\partial x^2} \right] \\ &= \frac{\partial f}{\partial t} + \langle u \rangle \frac{\partial f}{\partial x} - \nu_p \frac{\partial}{\partial x} \left(f \frac{\partial f}{\partial x} \right). \end{aligned}$$

Instead, attempts were made to use semigroup theory [14], representation of $\mathcal{L}$ as a positive kernel function, and placing constraints on the Fourier series solution [15, 16] to enforce preservation of positivity. However, a constructive, practical constraint on $\mathcal{L}$ could not be determined in the course of this work. Guaranteeing positivity of a Fourier series expansion of a positive function is an open area of inquiry, so the failure to derive constraints on $\mathcal{L}$ is likely an artifact of the basis functions for the problem rather than an inherent issue with the infinite-dimensional operator formulation. The failure to determine a means of enforcing positivity certainly does not mean that such a constraint cannot be found, but it does indicate that while several constraints were simple and intuitive to impose, other conditions will require more nuanced and complex approaches.

It should be noted that, if correctly parametrized, $\mathcal{L}$ would exactly represent the effects of dispersion on the evolution of $\langle c \rangle$. By definition,

$$\mathcal{L} \langle c \rangle = -\frac{\partial \langle u'c' \rangle}{\partial x}.$$

In terms of the Fourier series solution of $\langle c \rangle$, this equates to $\lambda_k \langle \hat{c}_k \rangle = -(ia_k) \left\langle \widehat{(u'c')}_k \right\rangle$. Solving for λ_k yields

$$\lambda_k = \frac{-(ia_k) \left\langle \widehat{(u'c')}_k \right\rangle}{\langle \hat{c}_k \rangle}.$$

Since both $\langle c \rangle$ and $\langle u'c' \rangle$ are functions of time, an exact parametrization would require time dependence in the eigenvalues λ_k. This time dependence cannot be recovered, however, because $\langle u'c' \rangle$ cannot be observed for the same reasons κ can't be observed. Even if $\mathcal{L}$ could exactly reproduce the effects of dispersion on $\langle c \rangle$, it would only be a descriptor of the contaminant's transport in the mean. It would not represent the effect of dispersion on the evolution of a single contaminant field, c, through a single heterogeneous porous medium. To provide a 95% confidence interval for the evolution of a single contaminant field, an additional representation of the evolution of its variance, $\langle c'^2 \rangle$, would be required. The uncertainty representation being developed here can thus be seen as a first-order descriptor of the uncertainty in the transport of the contaminant, while an uncertainty representation of the variance would be a second-order correction to the current formulation. First, the feasibility of developing a first-order representation must be established, which is the focus of this work.

Chapter 3

Bayesian inference of an uncertain operator

Initial work focused on assessing the feasibility of inferring the mean of the stochastic operator $\mathcal{L}$ in terms of its eigenvalues using observations of the evolution of the mean concentration only. This was done by attempting to infer the eigenvalues of the mean operator using Bayesian inference. How many eigenvalues could be inferred, and how precisely, was assessed based on frequency of observation and on whether observations were collected in a time series or across the spatial domain. Finally, the inferred eigenvalues were used to predict the evolution of $\langle c \rangle$ farther downstream or at later times than the inference data was collected. This was done to assess the predictive capability of the solution to the Bayesian inverse problem.

Bayesian inference of an operator has been explored previously. For instance, in the context of blind deconvolution, the kernel of a convolution operator is inferred along with an image [17], thereby recasting the problem of inferring an operator as a field inversion problem. Much work has focused on inference of the covariance matrix of a multivariate Gaussian distribution, for example to quantify uncertain measurement errors [18, 19]. These matrices are finite-dimensional and do not directly affect the dynamics of the state, only their presumed measurement error.

Non-Bayesian methods for inferring an operator from state observations have recently been developed. For instance, in [20] the operators in a reduced-order model are inferred

deterministically using data generated from a higher-fidelity model's output, taken at a variety of times, locations and model parameter values. The types of operators in question appear in dynamical systems and are often discretizations of differential operators, making their inference most similar to the proposed inference problem here, but inference is in a deterministic setting. This work instead focuses on the Bayesian inference of an infinite-dimensional differential operator's spectrum using observations of the state variable whose dynamics it affects.

Of particular interest is how the solution to the Bayesian inference problem depends on the amount and type of data, for example if observations are a spatial series or a time series, and how close the observations are to each other. The Bayesian inverse problem in terms of the mean operator's eigenvalues is posed by defining two distributions: the prior distribution, which represents the uncertainty in the eigenvalues prior to comparing to data, and the likelihood distribution, which represents the probability of observed data arising from the model in question for a specific set of eigenvalues. Generally speaking, the likelihood assigns low probability to a set of eigenvalues if it produces model outputs (in this case, predictions of $\langle c \rangle$) that have large discrepancies with the data. Defining the vector of eigenvalues to be $\mathbf{\Theta}$, the prior distribution as $p(\mathbf{\Theta})$, and the likelihood as $p(\mathbf{d}|\mathbf{\Theta})$, the solution of the Bayesian inverse problem is the posterior distribution, defined as

$$p(\mathbf{\Theta}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{\Theta})p(\mathbf{\Theta})}{\int p(\mathbf{d}|\mathbf{\Theta})p(\mathbf{\Theta})\mathrm{d}\mathbf{\Theta}}.$$

The denominator of this expression cannot be computed analytically because of the nonlinearity of the state with respect to the eigenvalues, which appears in the likelihood distributions. Instead, samples distributed according to the posterior distribution are generated using Markov Chain Monte Carlo (MCMC) [21].

Let the right-hand side of (2.1), denoted $\mathcal{D}$, be defined as

$$\mathcal{D}\left\langle c\right\rangle \equiv\left(\nu_{p}\frac{\partial^{2}}{\partial x^{2}}+\mathcal{L}\right)\left\langle c\right\rangle . \tag{3.1}$$

The uncertainty of $\mathcal{L}$ induces uncertainty in $\mathcal{D}$, whose eigenvalues are denoted μ_k. As discussed in Chapter 2, the eigenfunctions of $\mathcal{D}$ are the Fourier modes because of shift-invariance of the mean equations:

$$\mathcal{D}\exp\left(ia_{k}x\right)=\mu_{k}\exp\left(ia_{k}x\right).$$

Since λ_k can be determined from μ_k using the relation

$$\mu_{k}=-\nu_{p}a_{k}^{2}+\lambda_{k}, \tag{3.2}$$

inference of μ_k is equivalent to inference of λ_k.

Given the parametrization of the uncertain operator using its eigendecomposition (2.2), the goal is to infer the uncertain eigenvalues of $\mathcal{D}$, which forms the right-hand side of a generalized diffusion equation:

$$\begin{aligned}\frac{\partial\left\langle c\right\rangle }{\partial t}+\left\langle u\right\rangle \frac{\partial\left\langle c\right\rangle }{\partial x}&=\mathcal{D}\left\langle c\right\rangle ,\\ \left\langle c\right\rangle \left(0,t\right)&=\left\langle c\right\rangle \left(L_{x},t\right),\\ \left\langle c\right\rangle \left(x,0\right)&=c_{0}(x).\end{aligned} \tag{3.3}$$

The choice to perform inference in terms of $\mathcal{D}$ rather than $\mathcal{L}$ is because the requirement of a diffusive right-hand side is easier to impose on $\mathcal{D}$ directly, since it corresponds to $\Re\left[\mu_{k}\right]<0$.

First, to investigate the feasibility of inferring such an operator and the dependence of the Bayesian inference problem on the amount, type, and quality of data, a study is

performed with data generated using an operator whose eigenvalues are known *a priori*. The formulation of the inverse problem and the results of this study are presented in Section 3.4. Second, a Bayesian inverse problem is performed using data generated from a direct numerical averaging of the high-fidelity model defined in (1.1)-(1.3) exhibiting anomalous diffusion. The formulation of the Bayesian inverse problem and results of this study are presented in Section 3.5.

3.1 Prior specification

The prior distribution will be defined in terms of the real and imaginary parts of μ_k.Because $\langle c \rangle$ is real, its Fourier coefficients are conjugate symmetric; that is, $\langle \hat{c}_{-k} \rangle = \overline{\langle \hat{c}_k \rangle}$. As a result, the action of $\mathcal{D}$ on $\langle c \rangle$ can be expressed in terms of its positive wavenumber eigenvalues only:

$$\mathcal{D}\langle c \rangle = 2\,\Re\left[\sum_{k=0}^{\infty} \langle \hat{c}_k \rangle\, \mu_k \exp\left(i a_k x\right)\right].$$

Because of conservation of mass, $\mu_0 = 0$. The number of eigenvalues to be inferred is further limited, since the Fourier coefficients of $\langle c \rangle$ decay with respect to k and time. Because of this, the number of Fourier coefficients needed to fully resolve the Fourier series solution of (3.3) for any $t > 0$ will always be less than or equal to the number needed to resolve the initial condition $c_0(x)$, herein denoted N_k. Thus inference will focus on $\{\mu_k\}_{k=1}^{N_k}$.

The prior distributions for the real and imaginary parts of the eigenvalues of $\mathcal{D}$ are determined by considering the possible range of eigenvalues for fractional derivative operators ranging from the first to the second derivative. These operators are denoted ${}^{\partial^\alpha(\cdot)}/_{\partial x^\alpha}$ and their action is defined spectrally by ${}^{\partial^\alpha(e^{ia_k x})}/_{\partial x^\alpha} \equiv (ia_k)^\alpha e^{ia_k x}$. Non-integer fractional derivatives yield nonlocality in the solutions of PDEs in which they appear, which has motivated their use to represent anomalous diffusion [22]. Fractional derivatives with $\alpha \in [1, 2]$ are used here to determine plausible values for the eigenvalues of $\mathcal{D}$, whose effects lie somewhere between

pure advection ($\alpha = 1$) and pure diffusion ($\alpha = 2$). As shown in Figure 3.1, the real parts of the eigenvalues of fractional derivatives decrease monotonically as functions of the fractional power α for each k, while the imaginary parts achieve a maximum at varying fractional powers depending on k.

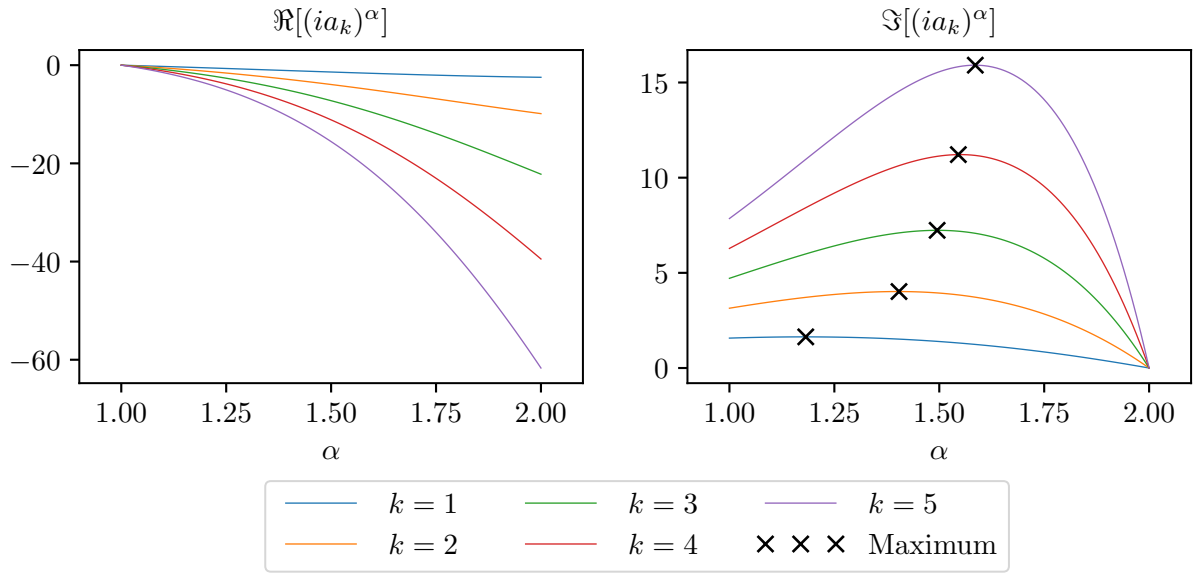

Figure 3.1: The real and imaginary parts of the eigenvalues of the fractional derivative with fractional power α varying from 1 to 2.

The real part of μ_k is constrained to be negative, so the negative real parts $-R_k$ are represented using an exponential distribution, $p(-R_k) = \exp(-R_k/\beta_k)/\beta_k$, in the prior. The scaling coefficients β_k are defined so that 95% of the probability mass for each R_k falls between the negative real parts of the eigenvalues for the first and second derivatives respectively, as shown in Figure 3.2. This is done using the CDF of $-R_k$, $P(-R_k) = 1 - \exp(-R_k/\beta_k)$:

$$0.95 = P(-(-\nu a_k^2)) = 1 - \exp\left(\nu a_k^2/\beta_k\right)$$

$$\Downarrow$$

$$\beta_k = \nu a_k^2 / \ln 0.05.$$

The ν used to define the prior was determined based on an argument that $\langle c \rangle$ should not decay completely after a single flowthrough (the time required to advect a domain length at the mean velocity, ${}^{L_x}/_{\langle u \rangle}$). Solutions with more rapid decay would diffuse away too quickly to provide enough information to inform the eigenvalues. To guarantee the solution has not decayed completely it is sufficient to guarantee that $|\langle \hat{c}_1 \rangle| \left(t = {}^{L_x}/_{\langle u \rangle}\right) \geq 10^{-12}$. Since the diffusion operator induces the most decay, it was assumed that $\mathcal{D} = \nu \partial_{xx}$ and the following inequality was solved for ν:

$$\begin{aligned} 10^{-12} \leq |\langle \hat{c}_k \rangle \left(t = {}^{L_x}/_{\langle u \rangle}\right)| &= \left| \langle \hat{c}_1 \rangle (0) \exp\left((-\nu a_k^2 - \langle u \rangle (i a_k)) \frac{L_x}{\langle u \rangle} \right) \right| \\ &= |\langle \hat{c}_1 \rangle (0)| \exp\left(-\nu a_k^2 \frac{L_x}{\langle u \rangle} \right). \end{aligned}$$

It was found that $\nu \approx 2.5$ was the largest diffusion coefficient for which the solution would not have decayed completely. This approximate maximum $\nu_{max} = 2.5$ was also used to define the upper bound on the imaginary prior distribution. It should be noted that this is more of a practical requirement than a true statement of prior knowledge. However, it is a loose constraint, yielding a broad, unbounded prior, so for informed directions the prior was dominated by the likelihood in the posterior distribution.

The negative real parts are defined on a domain bounded from below by zero. If the high-probability region of the posterior distribution is near this boundary, the mixing of the Markov chain will be poor. This is because MCMC samplers generally employ Gaussian proposal distributions, which will propose many samples outside the domain that are rejected. To improve mixing of the Markov chain, the natural logarithm of the negative real parts, denoted r_k, are inferred instead. Their prior distributions can be computed analytically

using a variable transformation and are defined as

$$\begin{aligned} r_k &= \ln(-R_k), \\ p(r_k) &= \frac{\exp\left(-e^{r_k}/\beta_k + r_k\right)}{\beta_k,} \end{aligned} \tag{3.4}$$

where β_k are the same as for the distributions of R_k. An example of this density is shown in Figure 3.2.

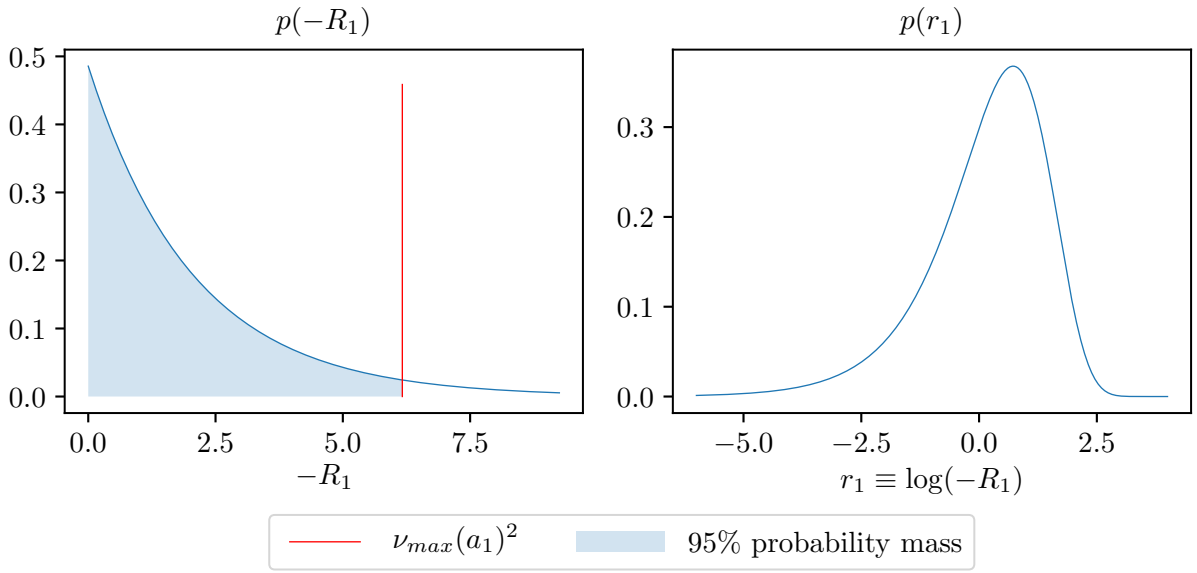

Figure 3.2: The prior distribution for R_1 and r_1.

The imaginary part is not bounded away from zero. The range of values considered probable are also determined by examining the imaginary part of the fractional derivative operator's eigenvalues. A normal distribution was used because it has infinite support, which guarantees it satisfies the conditions for the Bernstein-von Mises theorem (in the limit of infinite data the posterior becomes independent of the prior) [23]. The minimum imaginary part for fractional derivatives with $\alpha \in [1, 2]$ is known to be 0, from the second derivative operator. The maximum requires more care, since it varies as a function of k.

The imaginary part of the fractional derivative eigenvalue is

$$\Im[(ia_k)^\alpha] = (a_k)^\alpha \sin\left(\frac{\pi}{2}\alpha\right)$$

The α that maximizes this expression can be found using standard calculus techniques:

$$\begin{aligned} 0 &= \frac{\partial}{\partial\alpha}\left((a_k)^\alpha \sin\left(\frac{\pi}{2}\alpha\right)\right) \\ &= \alpha\,(a_k)^{\alpha-1}\sin\left(\frac{\pi}{2}\alpha\right) + \frac{\pi}{2}\,(a_k)^\alpha \cos\left(\frac{\pi}{2}\alpha\right). \end{aligned}$$

Then the α that maximizes the imaginary part for any k must satisfy

$$\alpha \tan\left(\frac{\pi}{2}\alpha\right) = -\frac{\pi}{2}a_k.$$

This cannot be solved analytically. Rather than perform a nonlinear solve for α for each k, an approximate maximum imaginary part, denoted m_k, was found by computing $\Im\left[(ia_k)^\alpha\right]$ for 100 values of α in the range $[1, 2]$ and taking the maximum of the set, as shown in Figure 3.1. The distribution for each I_k is defined so that 95% of the probability mass falls between 0 and $\nu_{max}\, m_k$, as shown in Figure 3.3:

$$p(I_k) = \mathcal{N}\left(\frac{\nu_{max}\, m_k}{2}, \left(\frac{\nu_{max}\, m_k}{4}\right)^2\right), \tag{3.5}$$

Inference is defined in terms of the parameters $\boldsymbol{\Theta} \equiv [r_1, r_2, \cdots, I_1, I_2, \cdots]$. All parameters are assumed independent, so the infinite-dimensional prior density would be defined as

$$p(\boldsymbol{\Theta}) \equiv \prod_{k=1}^{\infty} p(r_k)p(I_k).$$

Recall, however, that the problem is truncated to at most to $2N_k$, where N_k is the number of Fourier modes required to resolve the Fourier series of the initial condition of the state.

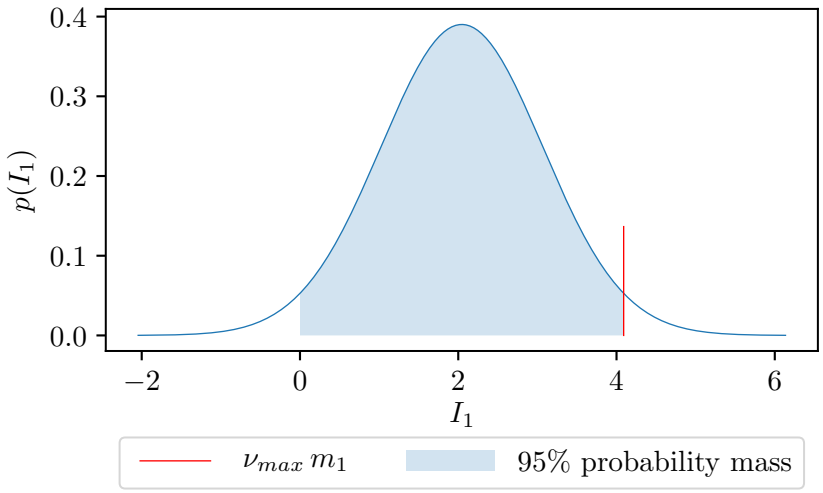

Figure 3.3: The prior distribution for I_1.

The truncated version of the prior for the inference problem is thus defined as

$$p(\mathbf{\Theta}) \equiv \prod_{k=1}^{K} p(r_k)p(I_k), \tag{3.6}$$

where $K \leq N_k$.

3.2 Likelihood specification

Several data sets will be considered, but they will exhibit additive, independent and normally-distributed measurement error. The data model is defined as

$$d_i = \langle c \rangle (x_i, t_i; \mathbf{\Theta}) + \epsilon_i, \quad i = 1, \cdots, N_{obs}, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2),$$

where σ_i may vary across observations. Defining the measurement covariance matrix Σ by $\Sigma_{ii} = \sigma_i^2$, $\Sigma_{ij} = 0, i \neq j$, the likelihood is

$$p(\mathbf{d}|\mathbf{\Theta}) = \frac{\exp\left(-\frac{1}{2}\left\|\mathbf{d} - \langle \mathbf{c} \rangle\right\|^2_{\Sigma^{-1/2}}\right)}{(2\pi)^{N_{obs}/2}|\Sigma|^{-1/2}}. \tag{3.7}$$

3.3 Inference implementation

Before inference a global variance-based sensitivity analysis is performed to determine to which eigenvalues $\langle c \rangle$ is most sensitive [24–26]. The sensitivity is assessed in terms of the Sobol total-effect index, a measure of the contribution to variance in $\langle c \rangle$ from varying an eigenvalue alone as well as from its variation along with other eigenvalues. The complete set of indices, one for each eigenvalue, sums to 1, with each index representing the fraction of variance in $\langle c \rangle$ that can be attributed to that eigenvalue. Any eigenvalues whose Sobol total-effect indices exceed 10^{-5} are included in the inference, and the rest are fixed at a reasonable value as described below. This analysis is performed using the Python software package SALib [27].

To generate samples of the posterior distributions of the eigenvalues using Markov Chain Monte Carlo (MCMC), the Delayed Rejection Adaptive Metropolis (DRAM) algorithm [28], implemented in the MIT UQ library MUQ [29], is used. The starting point of the Markov chain is determined by performing two deterministic optimizations, also using MUQ. The first optimization is performed with the assumption that $\mathcal{D}$ is of the form

$$\mathcal{D} = \nu \frac{\partial^\alpha}{\partial x^\alpha},$$

and ν, α are optimized to maximize the likelihood density. The second optimization relaxes the assumption on the form of $\mathcal{D}$ and maximizes the posterior density with respect to r_k and I_k and is started at the solution to the first optimization. Only the eigenvalues to which $\langle c \rangle$ is sensitive are optimized and included in the Bayesian inference. The insensitive eigenvalues are fixed at the solution to the first optimization. By fixing the insensitive eigenvalues at those of the fractional derivative, the Bayesian inference can be interpreted as finding a correction to the fractional derivative in terms of the sensitive eigenvalues.

DRAM generates proposed steps for the Markov chain using a Gaussian centered at the current position in parameter space. The implementation of DRAM in MUQ begins with an isotropic Gaussian for its proposal distribution, i.e. its covariance operator is of the form $s^2 I$, where s is a step size specified by the user. After a certain number of steps, this isotropic covariance operator is replaced with a sample covariance computed using the previous steps in the Markov chain. The sample covariance is then updated at regular intervals to continue improving the approximation of the shape of the posterior, thereby proposing fewer rejected steps.

DRAM using an initially isotropic proposal distribution can struggle to effectively explore the entire posterior parameter space if the extent of the high-probability region varies significantly in the different parameter directions. The step size must initially be small enough to produce accepted steps in the shortest direction, which means the step size is incredibly small for the longest direction. As a result, the sample covariance that is used to generate later steps after adapting the proposal distribution may be much smaller in the long directions than it is in reality, and will limit how these directions are explored. Theoretical results indicate that in the limit of infinite samples, the entire space will be explored and samples from the resulting Markov chain will be distributed according to the posterior, but in reality the number of steps necessary to reach this theoretical limit would be infeasibly large. To alleviate this issue, the priors for the imaginary parts I_k, which increase in extent with k, are normalized to $z_k \sim \mathcal{N}(0,1)$ for MCMC by

$$z_k = \frac{I_k - \mu_{I_k}}{\sigma_{I_k}} = \frac{I_k - \frac{m_k}{2}}{\frac{m_k}{4}}.$$

Corresponding samples of I_k are computed using the inverse transformation $I_k = z_k \sigma_{I_k} + \mu_{I_k}$.

3.3.1 Analysis of MCMC results

Length 2×10^5 chains were run for each data scenario. The first 1×10^5 samples were discarded as burn-in. The Kullback-Leibler divergence (DKL or KL divergence) is a natural measure of how much information was gained through inference [30], since it is a measure of how different two probability distributions are from each other. Positive KL divergence indicates information gain in the posterior compared to the prior, while a negative value indicates information loss. To measure the information gain in the real and imaginary part of each eigenvalue individually, the KL divergence between the marginal prior and posterior for each parameter Θ_k is computed for this work. Denoting the marginal prior $p(\Theta_k)$ and the marginal posterior $p(\Theta_k|\mathbf{d})$, the KL divergence is defined as

$$D\Big(p(\Theta_k|\mathbf{d}) \;\Big|\Big|\; p(\Theta_k)\Big) \equiv \int \ln\left(\frac{p(\Theta_k|\mathbf{d})}{p(\Theta_k)}\right) p(\Theta_k|\mathbf{d})\mathrm{d}\Theta_k.$$

If an analytical expression for the posterior were available, this integral could be approximated using Monte Carlo integration by

$$D\Big(p(\Theta_k|\mathbf{d}) \;\Big|\Big|\; p(\Theta_k)\Big) \approx \frac{1}{N_s}\sum_{i}^{N_s} \ln\Big[p\left(\Theta_k^{(i)} \,|\, \mathbf{d}\right)\Big] - \ln\Big[p\left(\Theta_k^{(i)}\right)\Big],\ \Theta_k^{(i)} \sim p(\Theta_k|\mathbf{d}),$$

where N_s is the number of samples used in the sample mean. Although an analytical expression is unavailable, one can be approximated using a Kernel-Density Estimate (KDE) approximation [31], built using samples from the posterior generated using MCMC. However, since samples of the posterior indicate a nearly-Gaussian posterior for this problem, Gaussian approximations of the marginal posteriors are made using the sample mean and variance of the Markov chain, rather than constructing KDEs for each one. The KL divergence between

a single parameter Θ_k's marginal posterior and prior will thus be approximated by

$$D\Big(p(\Theta_k|\mathbf{d})\,\Big|\Big|\,p(\Theta_i)\Big) \approx \frac{1}{N_s}\sum_{i=1}^{N_s} \ln\left[p_{GA}\left(\Theta_k^{(i)}\,\middle|\,\mathbf{d}\right)\right] - \ln\left[p\left(\Theta_k^{(i)}\right)\right],\ \Theta_k^{(i)} \sim p(\Theta_k|\mathbf{d}), \quad (3.8)$$

where p_{GA} denotes a Gaussian approximation.

3.4 Case 1: Data from Fractional Advection-Diffusion Equation

To study the success of Bayesian inference of $\mathcal{D}$ using observations of $\langle c\rangle$, initial studies focused on a case where it was known that $\mathcal{D}$ could exactly represent the underlying operator. Data was generated using a 1D fractional advection-diffusion equation (FRADE), evolving an initial Gaussian pulse:

$$\begin{aligned} \frac{\partial \langle c\rangle}{\partial t} + \langle u\rangle \frac{\partial \langle c\rangle}{\partial x} &= \nu \frac{\partial^\alpha \langle c\rangle}{\partial x^\alpha}, \quad x \in (0,4), \quad \alpha \in [1,2] \\ \langle c\rangle(0,t) &= \langle c\rangle(4,t), \\ \langle c\rangle(x,0) &= \exp\left(-\frac{(1-x)^2}{2(.1)^2}\right). \end{aligned} \quad (3.9)$$

Fractional PDEs can be seen as limiting forms to solutions of continuous-time-random-walk based models, which have gained popularity as models of anomalous diffusion through heterogeneous porous media [22]. An example of the time evolution of the concentration field generated from this model is shown in Figure 3.4. In this case it is known *a priori* that the true eigenvalues of $\mathcal{D}$ are $\mu_k = \nu(ia_k)^\alpha$. This makes it possible to study if the true values are recovered in different data scenarios.

3.4.1 Likelihood

Data was generated by evaluating the FRADE model over a range of times and locations. Random noise distributed according to $\mathcal{N}(0,\sigma^2)$ was added to the model evaluations to

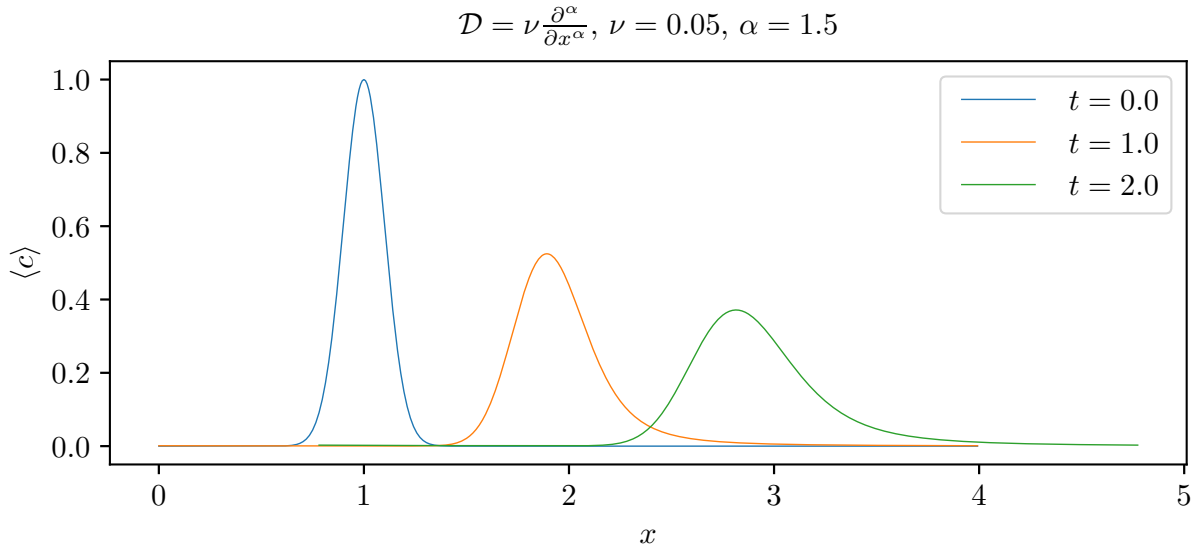

Figure 3.4: The evolution of a Gaussian initial condition with the FRADE at $t = 1$ and $t = 2$, 1/4 and 1/2 of a flowthrough time, respectively.

simulate measurement error. A measurement standard deviation of $\sigma = 0.005$, corresponding to a 1% standard error in the maximum concentration $\langle c \rangle = 1$, was used. Thus the likelihood is defined

$$p(\mathbf{d}|\mathbf{\Theta}) = \frac{1}{(2\pi\sigma^2)^{N_{obs}/2}} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{d} - \langle \mathbf{c} \rangle\|_2^2\right), \quad \sigma = 0.005,$$

where N_{obs} is the number of observations taken.

3.4.2 Results

The eigenvalues of $\mathcal{D}$ were inferred using spatial- and time-series data with 32, 64, or 512 observations. Observations were taken at regular intervals. For the spatial-series data, observations were taken at increasingly smaller intervals across the entire domain, as shown in Figure 3.5. For the time-series data, observations were also taken at increasingly smaller intervals from $t = 0$ to a specified final time. In both cases the observation window was

selected so that the entirety of the Gaussian pulse and its tails were observed.

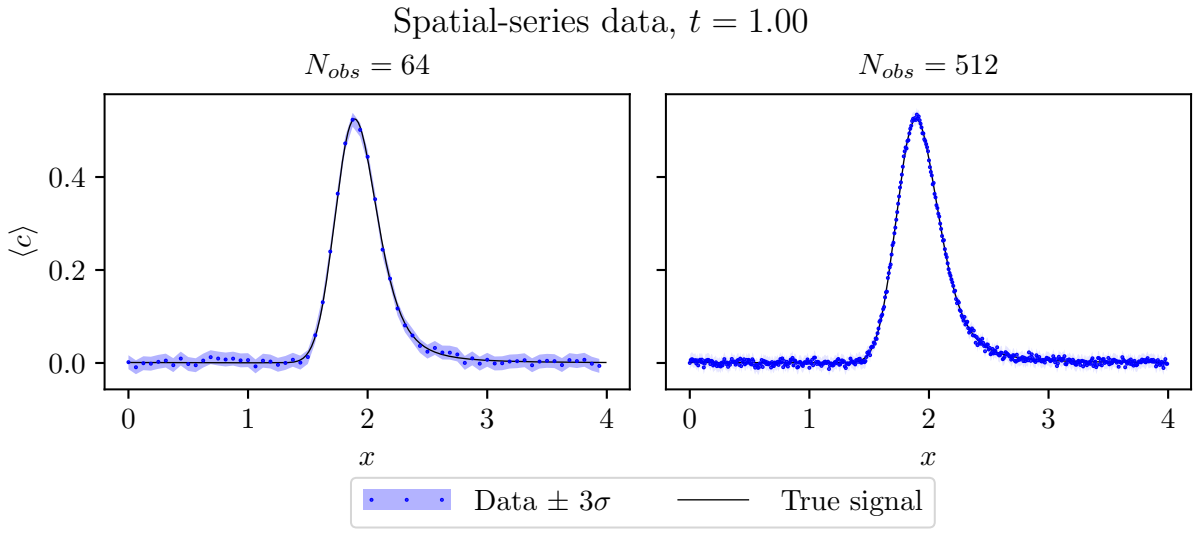

Figure 3.5: Spatial-series data taken over the entire domain, with increased frequency of observations corresponding to smaller intervals between observations.

Global variance-based sensitivity analysis was performed to determine how many eigenvalues to infer for each data scenario. The number of eigenvalues whose Sobol indices exceeded the tolerance of 10^{-5} for each scenario are reported in Table 3.1. The number of sensitive eigenvalues did not vary as a function of the number of observations taken for this study, so only the observation location or time for time-series or spatial-series data is reported. Unsurprisingly, the smoothing nature of diffusion in the problem limited the number of eigenvalues to which $\langle c \rangle$ was sensitive. Since only the eigenvalues that affect the evolution of $\langle c \rangle$ can be constrained by the likelihood, this limits the number of eigenvalues that can be inferred.

Cases with spatial data observed at a single time were sensitive to more eigenvalues than cases with time-series data observed at a single location. For cases with spatial-series data, the number of sensitive eigenvalues decreased as the time at which observations were taken increased. Since modes in the Fourier series solution decay as a function of time with higher modes decaying more rapidly, this is not surprising. The eigenvalues to which $\langle c \rangle$ was

Spatial series		**Time series**	
Observation time	# sensitive eigenvalues	Observation location	# sensitive eigenvalues
0.5	12	2.0	6
1.0	10	3.0	6
2.0	8	4.0	6

Table 3.1: The number of sensitive eigenvalues for varying data scenarios.

sensitive were, in general, well informed by the data. The measurement error for the problem is small, so this is expected. The maximum number of eigenvalues that were informed over all the cases considered was 12.

First, spatial-series data was used for inference. KL divergence plots were computed while varying the amount of data (see Figure 3.6) and the time at which data was collected (see Figure 3.7). Higher values of KL divergence indicate greater information gain. As shown in Figure 3.6, increased frequency of observation in the spatial domain led to more information gain in the eigenvalues that were informed by the data (see Figure 3.6).

The number of eigenvalues that were informed by the data depended on the time at which the spatial observations were made, as seen in Figure 3.7. For successively later times, the solution was sensitive to fewer and fewer eigenvalues. As seen for $t_{obs} = 2.0$, the real and imaginary parts for the highest wavenumber eigenvalue, μ_9, were not informed at all, indicated by a KL divergence ≈ 0. This is likely due to statistical error in approximating the Sobol indices for the sensitivity analysis, which caused μ_9 to be labeled sensitive when in reality it is not.

It is infeasible in realistic applications to have abundant spatial observations of the concentration field, since each observation location in the domain would require the drilling of a well. Instead it is more likely that one would have access to time-series observations of concentration at a limited number of locations. To reflect this, time-series data at one

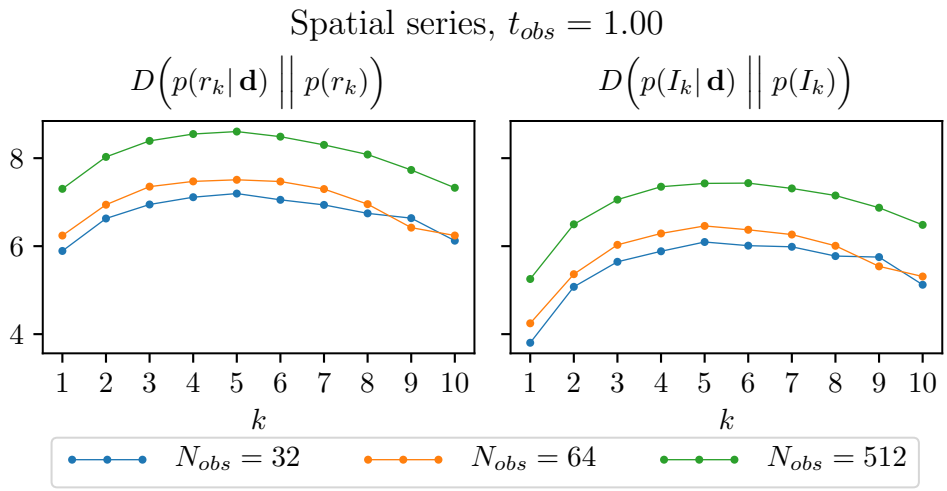

Figure 3.6: KL divergences for spatial-series data with varying frequencies of observation.

location was also used in the Bayesian inference of the eigenvalues.

For all the time-varying cases, $\langle c \rangle$ was deemed sensitive to only the first 6 eigenvalues. Once again, increased frequency of observation increased information gain for the informed eigenvalues, as shown in Figure 3.8. The information gain was less sensitive to the location at which the time-series data was collected than it was to the time at which spatial-series data was collected, as shown in Figure 3.9. This may be due to the fact that the observation locations were not far enough apart to provide unique information. A larger computational domain in the streamwise direction would allow for greater distance between observation locations and potentially unique information. While the number of eigenvalues that were informed by time-series data was less than by spatial-series data, the information gain in those that were informed is commensurate.

For both time-series and spatial data, the posterior distributions for cases with abundant ($N_{obs} = 512$) data contained the true value of the eigenvalues in their high-probability regions (see, e.g. Figure 3.10). Furthermore, the push-forward of the posterior to $\langle c \rangle$ outside the regime of the inference data was consistent with the true evolution, as shown in Figure 3.11. For the sparsest data ($N_{obs} = 32$), the posterior marginal distributions also largely contained

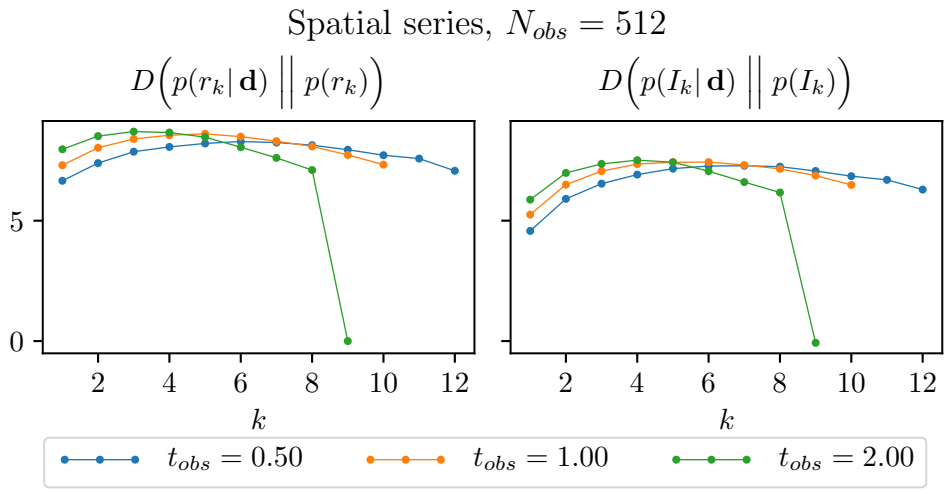

Figure 3.7: KL divergences for spatial-series data with varying observation time.

the true value of the eigenvalues in their high-probability regions (see, e.g. Figure 3.12). However, for some cases, the push-forward of the posterior outside of the regime of the inference data yielded nonphysical oscillations in the tails of $\langle c \rangle$ and negative concentrations, as shown in Figure 3.13. The likelihood can only penalize oscillations that induce large misfits with the data. Sparse observations allow for oscillations to occur between the data points. As shown in Figure 3.11, more frequent observations can ameliorate this issue. However, to guarantee that the inferred eigenvalues of the mean operator do not induce nonphysical evolutions, a constraint on the structure of $\mathcal{L}$ to preserve positivity would need to be derived. This constraint would also provide more prior information about the eigenvalues, which would further improve inference results.

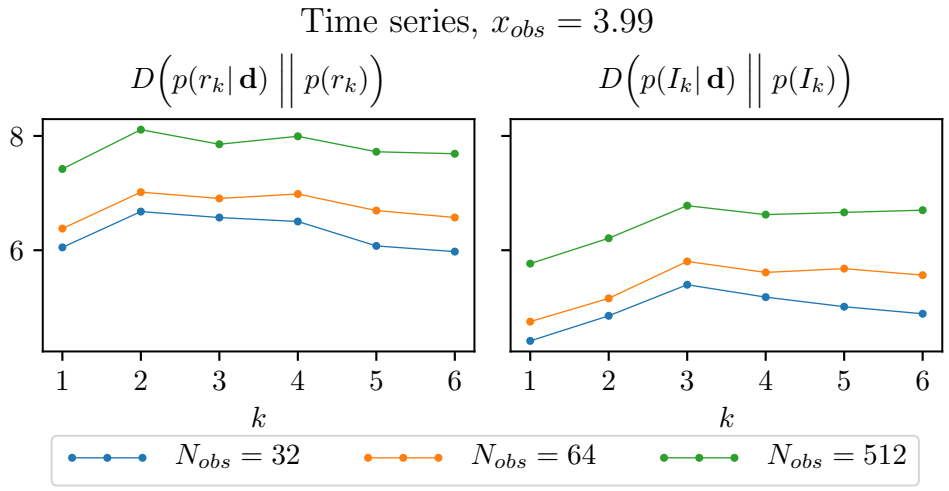

Figure 3.8: KL divergences for time-series data with varying frequencies of observation.

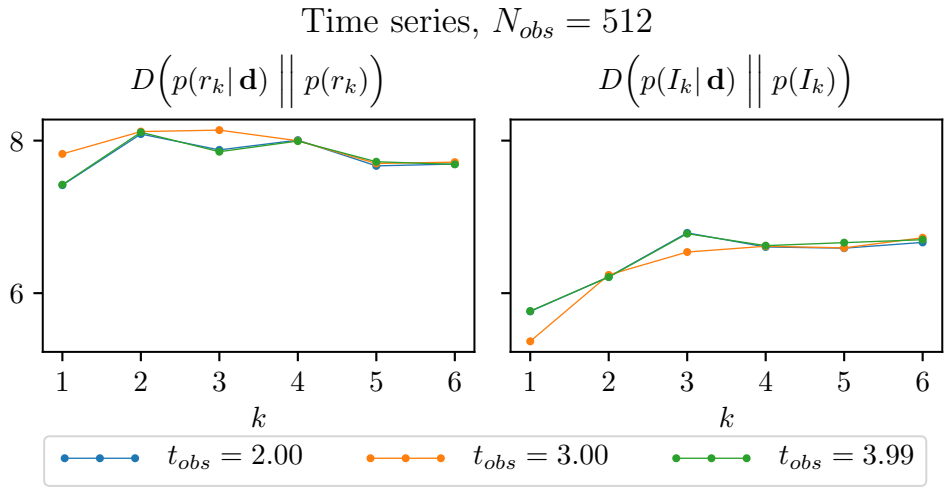

Figure 3.9: KL divergences for time-series data with varying observation location.

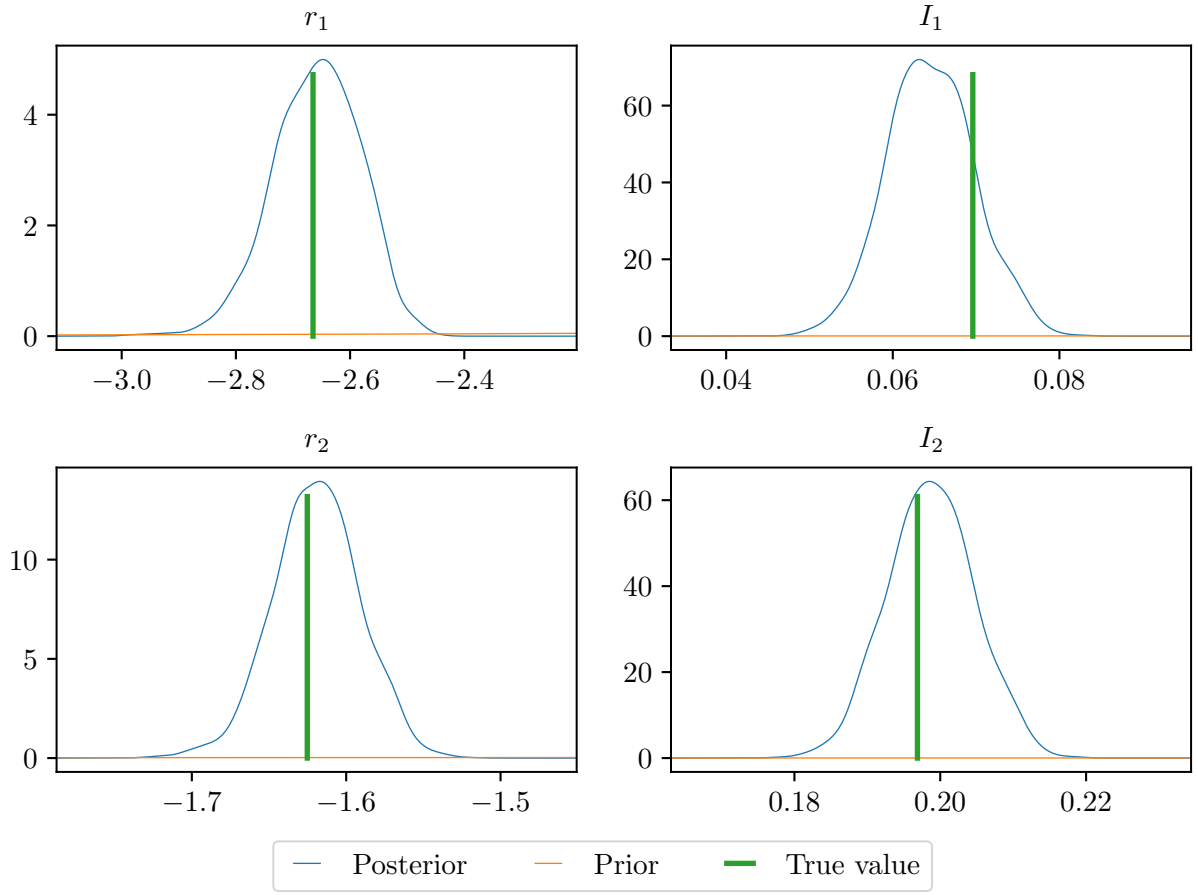

Figure 3.10: The posterior and prior marginal probabilities for the real and imaginary parts of the first two eigenvalues, inferred using 512 spatial observations taken at time $t = 0.5$.

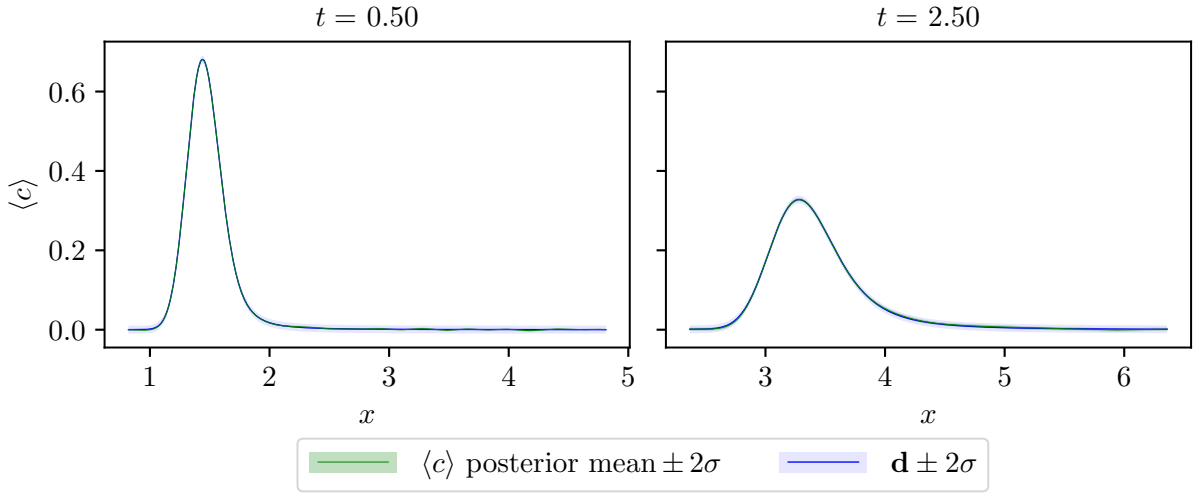

Figure 3.11: The push-forward of the posterior, inferred using 512 spatial observations collected at $t = 0.5$, to $\langle c \rangle$.

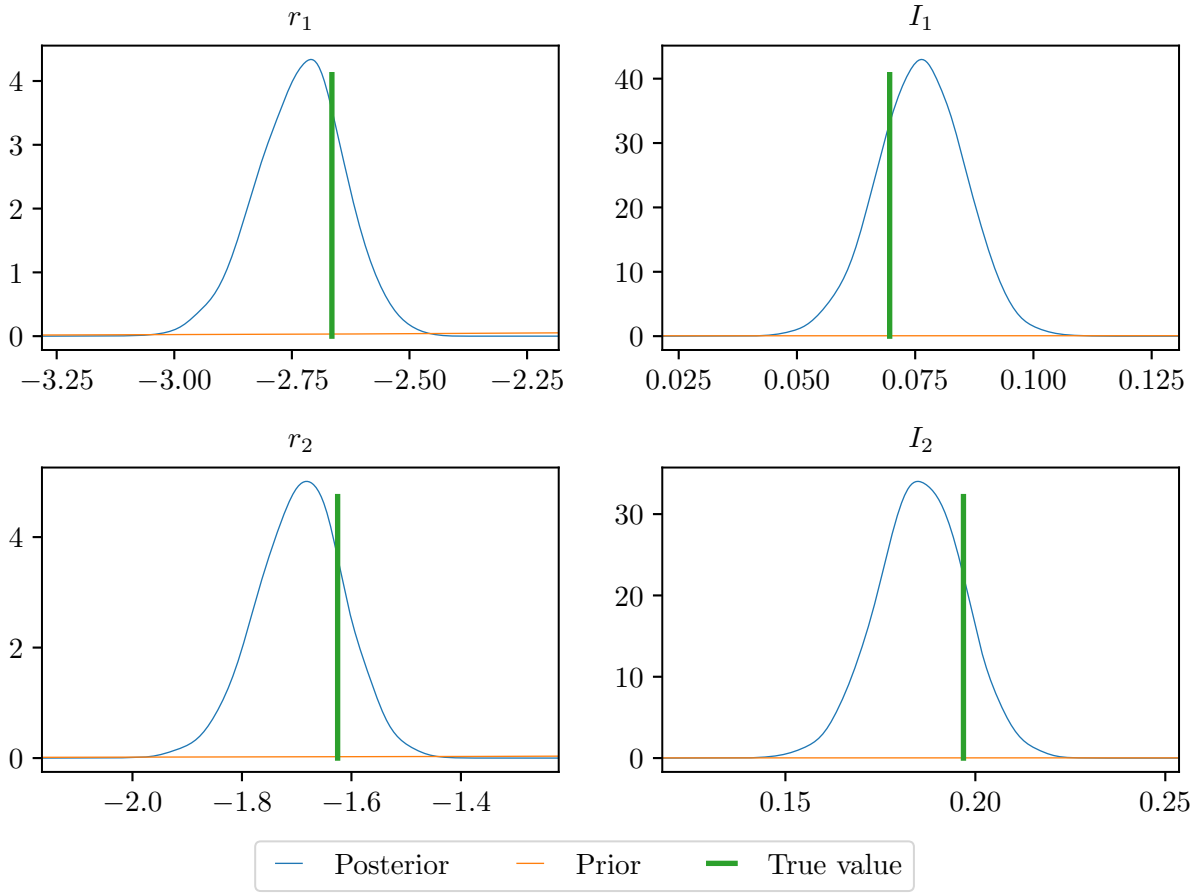

Figure 3.12: The posterior and prior marginal probabilities for the real and imaginary parts of the first two eigenvalues, inferred using 32 time-series observations taken at $x = 2.0$.

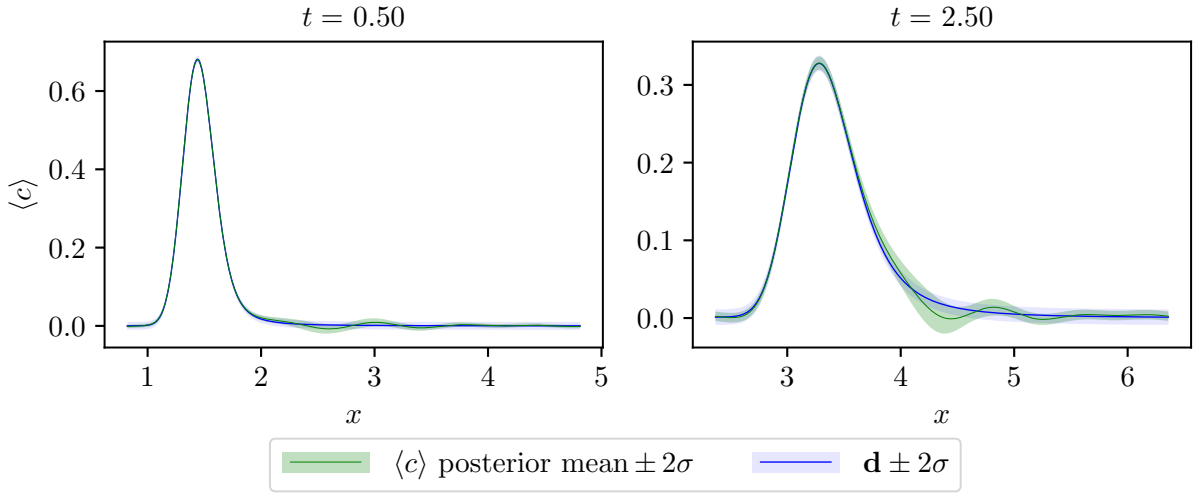

Figure 3.13: The push-forward of the posterior, inferred using 32 time-series observations collected at $x = 2.0$, to $\langle c \rangle$.

3.5 Case 2: Data from the direct numerical computation of $\langle c \rangle$

The inference in Section 3.4 using abundant data was largely successful in inferring the eigenvalues to which $\langle c \rangle$ was sensitive. Based on this success, inference was performed using data exhibiting anomalous diffusion, generated by computing an ensemble average of evolutions from the high-fidelity 2D advection-diffusion model (1.1)-(1.3). The goal of this inference was to determine if a mean $\mathcal{L}$ could be inferred from the high-fidelity data, and if it was valid outside of the regime of the inference data. To maximize information gain, spatial data collected at an early time, but still exhibiting anomalous diffusion, was used for the inference. The inferred eigenvalues were then used to predict $\langle c \rangle$ at later times, to study if it could successfully reproduce the evolution of $\langle c \rangle$.

3.5.1 Data and measurement error

The data used for the inference was collected from an ensemble average of depthwise-averaged solutions of the high-fidelity model defined in (1.1)-(1.3) with initial condition

$$c_0(x, y) = \exp\left(-\frac{(1-x)^2}{2(.1)^2}\right).$$

The ensemble average was computed over the space of permeability fields, which are assumed to be log-normally distributed:

$$\begin{aligned} \ln \kappa &\sim \mathcal{N}(0, C(\mathbf{x}, \mathbf{x}')), \\ C(\mathbf{x}, \mathbf{x}') &\equiv \sigma^2 \exp\left(-\frac{(x-x')^2}{2\ell_x^2} - \frac{(y-y')^2}{2\ell_y^2}\right). \end{aligned}$$

The assumption of log normality is common for models of heterogeneous porous media; see, e.g. [32, 33]. For the results presented here, the parameters used to define the permeability covariance operator are $\sigma^2 = 4.80603587627$, $\ell_x = 0.0955412170151$, and $\ell_y =$

0.0338019393384. This ensemble was generated as part of the study discussed in Chapter 4 and was chosen for this study because $\langle c \rangle$'s evolution exhibits anomalous diffusion. Sample velocities for the 2D ADE are generated by solving for the pressure as described in Appendix B and computing $\mathbf{u}^{(i)} = -\kappa^{(i)} \nabla p^{(i)}$. The implementation of the pressure solve is detailed in Appendix B.

Each velocity sample $\mathbf{u}^{(i)}$ was used to solve the 2D advection-diffusion equation (ADE), the implementation of which is described in Appendix A. The 2D evolution of c was averaged in y, and sample statistics were collected for $\langle c \rangle_y$ to approximate $\langle c \rangle$:

$$\langle c \rangle = \mathbb{E}\left[\langle c \rangle_y\right] \approx \frac{1}{N} \sum_{i=1}^{N} \langle c \rangle_y^{(i)} \equiv \langle c \rangle_N .$$

The sample variance was approximated using the Central Limit Theorem by

$$s_N^2 \approx \frac{1}{N-1} \left(\left\langle \langle c \rangle_y^2 \right\rangle_N - \langle c \rangle_{y,N}^2 \right),$$

computed pointwise within the domain.

The maximum possible variance in $\langle c \rangle$ is 1 because $c \in [0, 1]$. Assuming the Central Limit Theorem holds, $\langle c \rangle_N$ is a random variable with distribution

$$\langle c \rangle_N \sim \mathcal{N}\left(\langle c \rangle , \frac{\langle c'^2 \rangle}{N} \right).$$

A sample size of 576 was selected so that the variance of the sampling distribution would not exceed 5% of the maximum concentration $c = 1$, in the worst-case scenario where $\langle c'^2 \rangle = 1$. The data used for inference, shown in Figure 3.14, was a spatial sample taken at every point on the grid in the streamwise direction at time $t = 0.4$, or 1/10 of a flowthrough time ($L_x / \langle u \rangle = 4$).

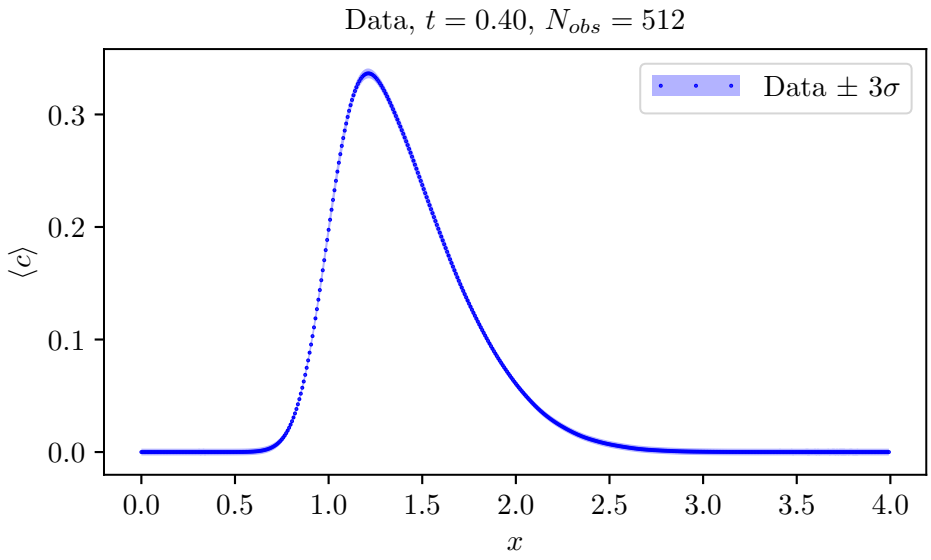

Figure 3.14: Even at 1/10 of a flowthrough, anomalous diffusion is already apparent using the detailed 2D advection-diffusion model.

3.5.2 Likelihood

The sample mean $\langle c \rangle_N$ was assumed to follow the Central Limit Theorem (CLT), so the data model was

$$d_i = \langle c \rangle_N (x_i) = \langle c \rangle (x_i) + \epsilon_i^{CLT}, \qquad \epsilon_{CLT} \sim \mathcal{N}(0, s_N^2(x_i)).$$

In the tails of the ensemble-averaged concentration field, the sample variance approached zero. A Dirac delta in the likelihood distribution would cause existing MCMC algorithms to struggle, so a minimum variance of 10^{-6} was assumed, yielding the likelihood

$$p(\mathbf{d}|\mathbf{\Theta}) = \exp\left(-\frac{1}{2} \left\| \langle \mathbf{c} \rangle (\mathbf{\Theta}) - \mathbf{d} \right\|_{\Sigma^{-1/2}}\right),$$

$$\Sigma_{ij} = \begin{cases} \max(s^2(x_i), 10^{-6}), & i = j, \\ 0, & i \neq j. \end{cases}$$

3.5.3 Results

The eigenvalues of $\mathcal{D}$ were inferred using abundant spatial data with low measurement error. The solution $\langle c \rangle$ from the generalized ADE was sensitive to the first 10 eigenvalues, whose KL divergences are shown in Figure 3.15. The resulting push-forward of the posterior parameter samples to data space indicates good agreement with the data used in the inference, as shown in Figure 3.16. However, by evolving the posterior parameter samples to later times as in Figure 3.17, it is clear that the inferred eigenvalues for $\mathcal{D}$ cannot successfully extrapolate in time. To reproduce the evolution of $\langle c \rangle$ the eigenvalues of $\mathcal{D}$ must be time-dependent. Since $\mathcal{D} \equiv \dfrac{\partial}{\partial x}\left(\nu_p \dfrac{\partial}{\partial x} + \mathcal{L}\right)$, the only possible source of this time dependence is $\mathcal{L}$. As discussed in Chapter 2, the eigenvalues of the deterministic operator $\mathcal{L}$ that would reproduce the effects of dispersion on the mean are time-dependent, so this does not come as a surprise.

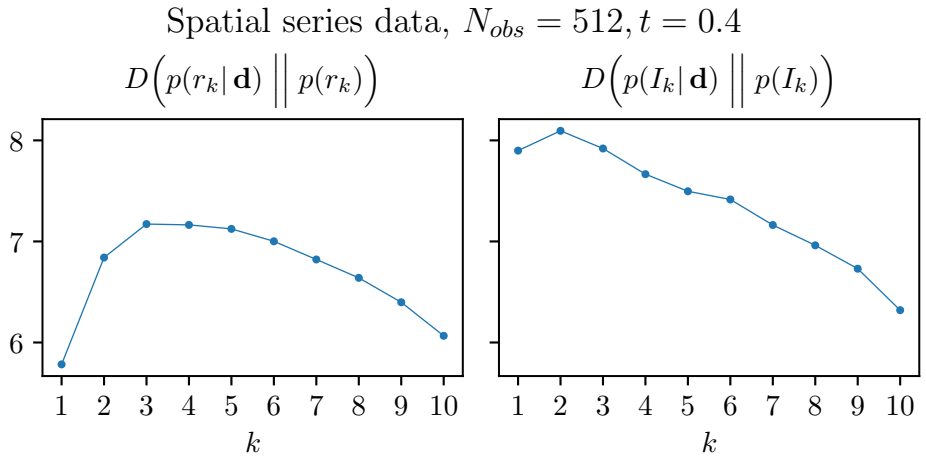

Figure 3.15: KL divergences for inference with data generated from the high-fidelity model.

3.6 Conclusions

The results in Section 3.4.2 and Section 3.5.3 indicate that, while it is possible to infer some of the eigenvalues of the mean of $\mathcal{L}$ using observations of $\langle c \rangle$, the diffusive nature of

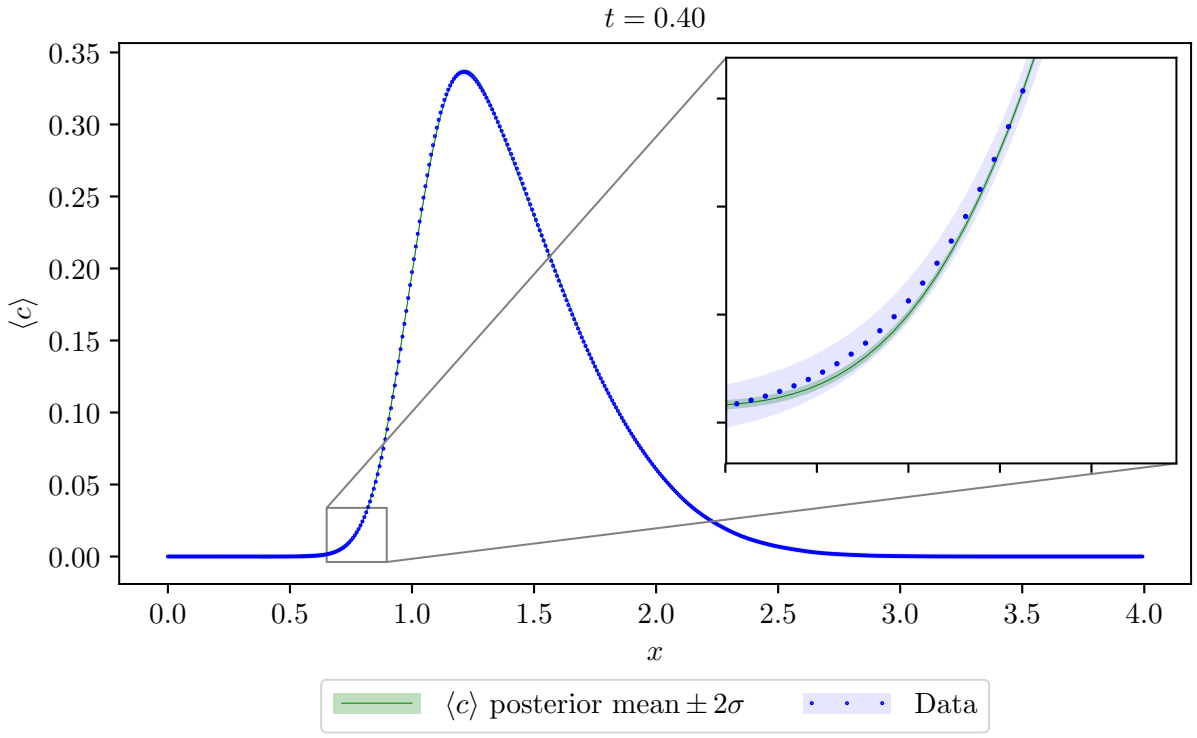

Figure 3.16: The push-forward of the posterior distribution to $\langle c \rangle$, compared to the inference data.

transport through porous media limits the number of eigenvalues for which this is possible. A goal of this work is to gain better understanding the nature of dispersion as a function of wavenumber and scenario by way of the spectrum of the stochastic operator $\mathcal{L}$. Because of this limitation in the number of eigenvalues that can be informed using observations of $\langle c \rangle$, another method of determining the spectrum of $\mathcal{L}$ is required. In Chapter 4, a novel method to directly compute eigenvalues at any desired wavenumber is detailed.

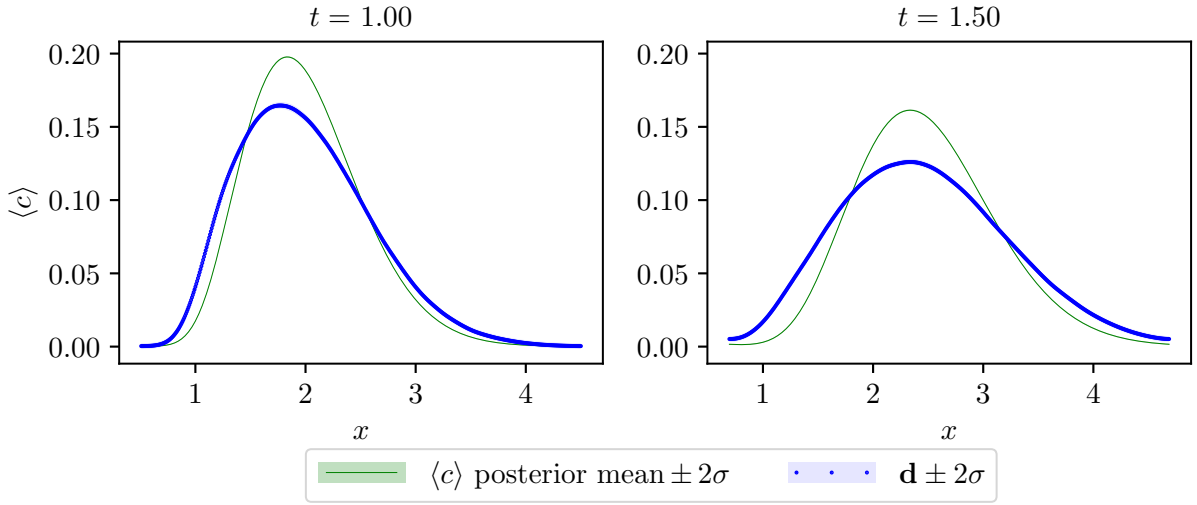

Figure 3.17: The push-forward of the posterior distribution to $\langle c \rangle$ at $t = 1.0$ and $t = 1.5$, an extrapolation from $t = 0.4$, the time the data used for inference was collected.

Chapter 4

Direct computation of the operator spectrum

As discussed in Chapter 3, the diffusive nature of transport through a porous medium limits the number of eigenvalues that can be informed using Bayesian inference based on observations of the mean evolution of a Gaussian pulse. Alternatively, this chapter describes a method to determine the eigenvalues using direct numerical simulation in a decoupled and highly-parallel fashion. Any statistical errors in the computed eigenvalues arising from this process can be quantified using classical frequentist statistical results such as the Central Limit Theorem.

Previous discussion has been in terms of the stochastic operator $\mathcal{L}$ acting on the mean state $\langle c \rangle$, defined such that

$$\mathcal{L}\langle c \rangle \equiv -\frac{\partial \langle u'c' \rangle}{\partial x},$$

where $(\cdot)' \equiv (\cdot) - \mathbb{E}\left[\langle \cdot \rangle_y\right]$. Instead the problem is recast in terms of a stochastic operator $\widetilde{\mathcal{L}}$ acting on the depthwise-averaged state $\langle c \rangle_y$. Because it is defined to act on $\langle c \rangle_y$ instead of the statistical mean, the assumption of shift-invariance does not extend to $\widetilde{\mathcal{L}}$. However, $\langle c \rangle_y$ is periodic in the simulations performed here, so the modeling ansatz was made that the eigenfunctions of $\widetilde{\mathcal{L}}$ are the Fourier modes. Similar to $\mathcal{L}$, its eigenvalues $\widetilde{\boldsymbol{\lambda}} = \left[\widetilde{\lambda}_1, \widetilde{\lambda}_2, \cdots\right]$ are stochastic. For a single velocity $\mathbf{u}^{(i)}$ sampled from the velocity distribution, there is a

corresponding operator $\widetilde{\mathcal{L}}^{(i)}$ such that

$$\widetilde{\mathcal{L}}^{(i)} \langle c \rangle_y^{(i)} \equiv -\frac{\partial \left\langle u'^y c'^y \right\rangle_y^{(i)}}{\partial x}.$$

Here $(\cdot)'^y \equiv (\cdot) - \langle \cdot \rangle_y$ denotes a deviation from the depthwise average only. Equivalently, its eigenvalues $\widetilde{\lambda}_k^{(i)}$ are defined by the relationship

$$\widetilde{\lambda}_k^{(i)} \langle \hat{c}_k \rangle_y^{(i)} \equiv -(ia_k) \left\langle \widehat{(u'^y c'^y)}_k \right\rangle_y^{(i)}. \tag{4.1}$$

Since the eigenvalues are derived from the dispersion, their distribution $p(\widetilde{\boldsymbol{\lambda}})$ depends on the statistics of the velocity $\mathbf{u}$. The dispersion is computed by evolving the high-fidelity, 2D detailed advection-diffusion equation, whose implementation is detailed in Appendix A. The detailed state c and $\mathbf{u}^{(i)}$ are then used to compute the depthwise-averaged $\left\langle u'^y c'^y \right\rangle_y$.

The connection between $\widetilde{\mathcal{L}}$ and $\mathcal{L}$ is best seen in terms of $\langle u'c' \rangle$ and $\left\langle u'^y c'^y \right\rangle_y$. Let $\mathbf{u} = [u, v]$, and note that the depthwise average of u is a constant due to mass conservation:

$$0 = \left\langle \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right\rangle_y = \frac{\partial \langle u \rangle_y}{\partial x} + \frac{\partial \langle v \rangle_y}{\partial y} = \frac{\partial \langle u \rangle_y}{\partial x}.$$

Then the depthwise average of u is equivalent to the bulk (domain-averaged) velocity. It is possible for the bulk velocity to vary statistically across samples $u^{(i)}$. However, the goal of this work is to account for the unobservable microstructural dependence on the velocity's local fluctuations from the mean. Because of this, and because the effects of statistical variation of the bulk velocity can easily be treated separately, the distribution of the velocity fields is constrained to have a constant deterministic bulk velocity, $\langle u \rangle$. Then taking the

expectation of $\left\langle u'^{y} c'^{y} \right\rangle_y$ yields

$$\begin{aligned}
\mathbb{E}\left[\left\langle u'^{y} c'^{y} \right\rangle_y\right] &= \mathbb{E}\left[\left\langle \left(u - \langle u \rangle_y\right)\left(c - \langle c \rangle_y\right)\right\rangle_y\right] = \mathbb{E}\left[\left\langle \left(u - \langle u \rangle\right)\left(c - \langle c \rangle_y\right)\right\rangle_y\right] \\
&= \mathbb{E}\left[\langle uc \rangle_y - \langle u \rangle \langle c \rangle_y - \langle u \rangle \langle c \rangle_y + \langle u \rangle \langle c \rangle_y\right] \\
&= \langle uc \rangle - \langle u \rangle \langle c \rangle = \langle u'c' \rangle \,.
\end{aligned}$$

Then $\mathbb{E}\left[\widetilde{\mathcal{L}}\,\langle c \rangle_y\right] = \mathcal{L}\,\langle c \rangle$, or, in terms of the eigenvalues $\widetilde{\lambda}_k$ and λ_k,

$$\lambda_k \langle \hat{c}_k \rangle = \mathbb{E}\big[\widetilde{\lambda}_k \langle \hat{c}_k \rangle_y\big] = \mathbb{E}\big[\widetilde{\lambda}_k\big] \langle \hat{c}_k \rangle + \mathbb{E}\big[\widetilde{\lambda}_k'^{\mathbb{E}} \langle \hat{c}_k \rangle_y'^{\mathbb{E}}\big], \tag{4.2}$$

where $(\cdot)'^{\mathbb{E}}$ denotes deviation from the statistical mean. The first term $\mathbb{E}\big[\widetilde{\lambda}_k\big] \langle \hat{c}_k \rangle$ encapsulates the mean effects of dispersion on the mean concentration. The second term $\mathbb{E}\big[\widetilde{\lambda}_k'^{\mathbb{E}} \langle \hat{c}_k \rangle_y'^{\mathbb{E}}\big]$ describes how the eigenvalues and Fourier coefficients covary, so it can be understood to encapsulate the effects of statistical variations in the dispersion on the mean.

For each sample $\widetilde{\lambda}_k^{(i)}$, a corresponding sample Fourier coefficient $\langle \hat{c}_k \rangle_y^{(i)}$ is defined through the depthwise-averaged evolution equation of the k^{th} Fourier coefficient:

$$\frac{\partial \langle \hat{c}_k \rangle_y^{(i)}}{\partial t} + i(a_k) \langle u \rangle \langle \hat{c}_k \rangle_y^{(i)} = -\nu_p^2 a_k^2 \langle \hat{c}_k \rangle_y^{(i)} + \widetilde{\lambda}_k^{(i)} \langle \hat{c}_k \rangle_y^{(i)} \,. \tag{4.3}$$

If $p\big(\widetilde{\boldsymbol{\lambda}}\big)$ were known exactly, the derived distribution of each $\langle \hat{c}_k \rangle_y^{(i)}$ would be specified through (4.3) and $\mathbb{E}\big[\widetilde{\lambda}_k \langle \hat{c}_k \rangle\big]$ could be computed directly. The effect of dispersion on the evolution of the mean $\langle c \rangle$ would be captured perfectly. However, as seen in the relation (4.1), $\widetilde{\lambda}_k$ once again depends on the unobservable evolution of the microstate. Then the goal is to develop models for the dependence of $p\big(\widetilde{\boldsymbol{\lambda}}\big)$ on summary statistics of the velocity that can be known *a priori*, such as variances and correlation lengths. These are only coarse descriptors of the microstructural behavior, so uncertainties will remain in the model for $p\big(\widetilde{\boldsymbol{\lambda}}\big)$. These

uncertainties are addressed in the stochastic formulation of $\widetilde{\mathcal{L}}$, discussed in Chapter 5.

The nature of the dependence of $p(\widetilde{\boldsymbol{\lambda}})$ on the statistics of $\mathbf{u}$ is probed by generating velocity ensembles with different variances and correlation lengths, with the corresponding evolutions of $\langle c \rangle$ exhibiting varying degrees of anomalous diffusion. Each ensemble of velocities is used to generate a corresponding ensemble of eigenvalues $\widetilde{\lambda}_k$ using (4.1). Summary statistics such as sample means and covariances of the eigenvalues are studied to determine their dependence on the statistics of $\mathbf{u}$. These observations are encapsulated in the stochastic formulation of $\widetilde{\mathcal{L}}$ in Chapter 5.

The remainder of this chapter focuses on the computation and analysis of eigenvalue ensembles across a variety of velocity statistics. The process of computing ensembles of velocities and corresponding eigenvalues is described in Section 4.1; determining a representative collection of scenarios based on velocity statistics is described in Section 4.2; and analysis of the distributions $p(\widetilde{\boldsymbol{\lambda}})$ across these scenarios is presented in Section 4.3.

4.1 Generating eigenvalue ensembles

As mentioned above, generating an ensemble of eigenvalues first requires the generation of an ensemble of velocities. Each velocity in the ensemble is used to evolve the high-fidelity 2D ADE, and the time-history of the solution is used to compute the eigenvalue. The ensemble of velocities is computed using an ensemble of permeability fields and Darcy's law. Generation of these ensembles is described in the order they occur: in Section 4.1.1, generation of a permeability field ensemble is described; in Section 4.1.2, the corresponding velocity ensemble; and in Section 4.1.3 the corresponding eigenvalue ensembles.

4.1.1 Generating an ensemble of permeability fields

As is commonly done [32, 33], the permeability fields in this work are assumed to be distributed log-normally, i.e. $\ln \kappa \sim \mathcal{N}(0, C)$. Let $\mathbf{x} = [x, y]$. The covariance operator C is defined as

$$C(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{1}{2}\left(\frac{(x-x')^2}{\ell_x^2} + \frac{(y-y')^2}{\ell_y^2}\right)\right),$$

where ℓ_x is the streamwise correlation length, ℓ_y is the depthwise correlation length, and σ^2 is the zero-separation variance.

As described in [34, 35], samples of the zero-mean random field $\ln \kappa$ can be generated using a Karhunen-Loeve (KL) expansion, defined as

$$\ln \kappa = \sum_{i=1}^{\infty} \sqrt{\gamma_i} \xi_i e_i(\mathbf{x}),$$

where $\{\gamma_i, e_i\}$ are the eigenvalues and eigenfunctions of the covariance operator C and ξ_i are scalar random variables. The covariance operator is shift invariant, so its eigenfunctions are known to be the Fourier modes. Its eigenvalues are thus its Fourier coefficients. Here they are computed by evaluating C on a regularly-spaced grid and performing a Discrete Fourier Transform (DFT). The structure of C and thus of its Fourier coefficients are dependent on the parameters ℓ_x, ℓ_y, and σ^2. In this case, the KL expansion of $\ln \kappa$ is equivalent to a Fourier expansion with random coefficients ξ_i.

Because $\ln \kappa$ is normally distributed, it is known that the scalar random variables ξ_i are distributed according to $\mathcal{N}(0, 1)$ [34]. To maintain consistency with a Fourier representation of $\ln \kappa$, ξ_i are assumed to be complex-valued standard normal random variables, i.e. $\Re[\xi_i], \Im[\xi_i] \sim \mathcal{N}(0, 1)$. Thus a sample of the log permeability can be obtained by drawing samples for ξ_{jk}, $j = 1, \cdots, N_x - 1$, $k = 1, \cdots, N_y - 1$ and computing the inverse

DFT:

$$\ln \kappa = \sum_{j=1}^{N_x-1} \sum_{k=1}^{N_y-1} \sqrt{\gamma_{jk}} \xi_{jk} \exp\left(i \left(\frac{2\pi j}{L_x} x_j + \frac{2\pi k}{L_y} y_k \right) \right),$$

where γ_{jk} are the Fourier coefficients of C.

While the permeability is constrained to be periodic in x, no such constraint is imposed in the depthwise direction. However, the speed of the DFT makes the sampling strategy above desirable. Thus a periodic field in x and y is sampled on the domain $[0, L_x] \times [-L_y, 2L_y]$, then truncated to the desired computational domain $[0, L_x] \times [0, L_y]$. Because the depthwise correlation length ℓ_y is constrained to be less than L_y, padding the periodic domain by L_y at the top and bottom guarantees the correlation between points at $y = 0$ and $y = L_y$ is negligible. Finally, the truncated field is exponentiated to obtain a random permeability κ with the desired covariance structure.

4.1.2 Generating an ensemble of velocities

A velocity sample is computed from a permeability field by solving the system

$$\begin{aligned} \mathbf{u} &= -\kappa \nabla p, \\ \nabla \cdot (\mathbf{u}) &= 0, \\ \mathbf{u}(0, y) &= \mathbf{u}(L_x, y), \\ v(x, 0) &= v(x, L_y) = 0. \end{aligned}$$

This amounts to solving for the pressure via a variable-coefficient Poisson problem

$$\begin{aligned}
\nabla \cdot (\kappa \nabla p) &= 0, \\
\frac{\partial p}{\partial y} &= 0, \quad y = 0, L_y, \\
\left(\frac{\partial p}{\partial x}\right)_{x=0} &= \left(\frac{\partial p}{\partial x}\right)_{x=L_x}, \\
\left(\frac{\partial p}{\partial y}\right)_{x=0} &= \left(\frac{\partial p}{\partial y}\right)_{x=L_x},
\end{aligned}$$

and computing $\mathbf{u} = -\kappa \nabla p$. Details of the implementation of the pressure solve are reported in Appendix B. The pressure solve is implemented using Finite Elements, while the advection-diffusion solver is implemented using a Fourier-Galerkin B-Spline-collocation scheme. The velocity must be projected onto the Fourier B-Spline representation for the advection-diffusion solver, but differences between the discrete differential operators causes the interpolated velocity to violate discrete continuity. To alleviate this issue a divergence-free projection is performed as detailed in Appendix C.

Finally, the computed velocities are scaled by the bulk streamwise velocity, thereby guaranteeing that all bulk velocities and thus the mean streamwise velocity of the ensemble are equal to 1. This rescaling is equivalent to modifying the driving streamwise pressure differential across the domain. Such a rescaling ensures that the effects of dispersion are distinguished from the effects of bulk velocity fluctuation.

4.1.3 Computing $\widetilde{\lambda}_k$ samples

For a given sample from the velocity ensemble, $\mathbf{u}^{(i)}$, the corresponding sample eigenvalue $\widetilde{\lambda}_k^{(i)}$ is defined such that

$$\widetilde{\lambda}_k^{(i)} \left\langle \hat{c}_k \right\rangle_y^{(i)} \equiv -(ia_k) \left\langle \left(\widehat{u'^y c'^y} \right)_k \right\rangle_y^{(i)},$$

where k is the index of the corresponding Fourier mode and $a_k = 2\pi k/L_x$ is the wavenumber. The terms $\langle \hat{c}_k \rangle_y^{(i)}$ and $\left\langle (\widehat{u'^y c'^y})_k \right\rangle_y^{(i)}$ can be computed by taking the depthwise average of a time-history evolved using the high-fidelity 2D ADE.

To isolate the dispersion's behavior to a single wavelength, a single Fourier mode is used as the initial condition for the detailed model. Unlike the mean evolution equation, for which the Fourier modes are eigenfunctions, other Fourier modes will be excited as the detailed system is evolved and fluctuations develop in the microstate. Since the ADE preserves mass, this means mass passes from the wavelength of the initially-excited Fourier mode to other wavelengths. To counteract decay in the mode of interest, a forcing function is introduced that maintains the norm of its corresponding coefficient. Each Fourier mode evolves at different a timescale, with lower frequency modes evolving more slowly than higher frequency modes, which require a finer time resolution. Evolving multiple modes simultaneously would require the runtime needed for the lowest-frequency mode, but the smallest timestep needed for the highest-frequency mode. By observing the evolution of each mode individually, time discretizations and runtimes can be scaled appropriately, which reduces overall cost.

To generate a sample eigenvalue $\widetilde{\lambda}_k^{(i)}$, the high-fidelity 2D ADE is evolved with velocity sample $\mathbf{u}^{(i)}$ and the initial condition set to its corresponding eigenfunction, the Fourier mode $\exp(ia_k x)$, where $a_k = 2\pi k/L_x$:

$$\frac{\partial c}{\partial t} + \mathbf{u}^{(i)} \cdot \nabla c = \nu_p \Delta c + f, \qquad x \in (0, L_x), \quad y \in (0, L_y) \tag{4.4}$$

$$c(0, y) = c(L_x, y), \qquad y \in (0, L_y) \tag{4.5}$$

$$\frac{\partial c}{\partial y} = 0, \qquad y = 0, L_y, \quad x \in [0, L_x] \tag{4.6}$$

$$c_0(x, y) = \exp(ia_k x). \tag{4.7}$$

The domain is nondimensionalized with respect to the depthwise domain length, so that $L_y = 1$ and $L_x = 4L_y$. The forcing function f is chosen to be of the form $f = \alpha(t) \langle \hat{c}_k \rangle_y$

where $\alpha \in \mathbb{R}$ so that it does not affect the phase of $\langle \hat{c}_k \rangle$. The numerical solution of (4.5)-(4.7) is detailed in Appendix A.

The scaling α is determined by requiring that $|\langle \hat{c}_k \rangle_y| = 1 \ \forall t$, which is equivalent to requiring $\partial_t |\langle \hat{c}_k \rangle_y|^2 = 0 \ \forall t$. A manipulation of (4.5), detailed in Section A.2.1, yields

$$\alpha \equiv \frac{-\Re\left[\langle \hat{c}_k \rangle_y^* \left(-\nu_p(a_k)^2 \langle \hat{c}_k \rangle_y - (ia_k) \left\langle \widehat{(\mathbf{u}c)}_k \right\rangle_y\right)\right]}{\left|\langle \hat{c}_k \rangle_y\right|^2}.$$

The physical meaning of the forcing as well as the real and imaginary parts of $\widetilde{\lambda}_k$ can be understood by inspecting the depthwise-averaged advection-diffusion equation in wavespace, where the sample index $(\cdot)^{(i)}$ is dropped for simplicity:

$$\frac{\partial \langle \hat{c}_k \rangle_y}{\partial t} + \langle u \rangle_y (ia_k) \langle \hat{c}_k \rangle_y = -\nu_p(a_k)^2 \langle \hat{c}_k \rangle_y - (ia_k) \left\langle \widehat{(u'^y c'^y)}_k \right\rangle_y + \alpha \langle \hat{c}_k \rangle_y .$$

Substituting $\widetilde{\lambda}_k \langle \hat{c}_k \rangle_y$ for $-(ia_k) \left\langle \widehat{(u'^y c'^y)}_k \right\rangle_y$ yields

$$\frac{\partial \langle \hat{c}_k \rangle_y}{\partial t} + \langle u \rangle_y (ia_k) \langle \hat{c}_k \rangle_y = -\nu_p(a_k)^2 \langle \hat{c}_k \rangle_y + \widetilde{\lambda}_k \langle \hat{c}_k \rangle_y + \alpha \langle \hat{c}_k \rangle_y .$$

Introducing the polar form of $\langle \hat{c}_k \rangle_y = r_k e^{i\theta_k}$ into the above equation and noting that forcing ensures $r_k = 1 \ \ \forall t$, the equation becomes

$$i\frac{\partial \theta_k}{\partial t} e^{i\theta_k} + \langle u \rangle_y (ia_k) e^{i\theta_k} = -\nu_p(a_k)^2 e^{i\theta_k} + \widetilde{\lambda}_k e^{i\theta_k} + \alpha e^{i\theta_k}.$$

Finally, solving for $\widetilde{\lambda}_k$ yields

$$\widetilde{\lambda}_k = \left(-\alpha - [-\nu_p(a_k)^2] \right) + i \left(\frac{\mathrm{d}\theta_k}{\mathrm{d}t} - \left[-(a_k) \langle u \rangle_y \right] \right).$$

The scaling term α counteracts decay in the coefficient $\langle \hat{c}_k \rangle_y$, so $-\alpha$ can be seen as the total diffusion in the system. The real part of $\widetilde{\lambda}_k$ is the total diffusion minus the contribution from pore-scale diffusion, leaving the contribution to diffusion from dispersion. Similarly, the imaginary part of $\widetilde{\lambda}_k$ is the rate of change in the phase of the coefficient that is not attributed to bulk advection, but rather to dispersion.

A sample eigenvalue $\widetilde{\lambda}_k^{(i)}$ can thus be computed using y-averaged quantities computed from the high-fidelity model output in one of two equivalent ways:

$$\widetilde{\lambda}^{(i)}(t) = \left(-\alpha^{(i)}(t) - [-\nu_p (a_k)^2] \right) + i \left(\frac{\mathrm{d}\theta_k^{(i)}}{\mathrm{d}t} - \left[-(a_k) \langle u \rangle_y \right] \right)$$

or

$$\widetilde{\lambda}^{(i)}(t) = \left\langle (\widehat{u'^y c'^y})_k \right\rangle_y^{(i)} = (\langle \widehat{u'^y c'^y} \rangle_y)_k^{(i)} = (\widehat{\langle uc \rangle}_y)_k^{(i)} - \langle u \rangle_y \langle \hat{c}_k \rangle_y^{(i)} ,$$

where $\langle u \rangle_y$ is not indexed by sample because it is always 1. Sample evolutions from the high-fidelity model can be computed in parallel, both in terms of the velocity ensemble and in terms of the Fourier mode initial conditions. Velocity ensembles are generated offline, and all samples are depthwise averaged. Sample statistics are computed in postprocessing as detailed in Appendix D. Additional samples can be generated if Central Limit Theorem estimates of sampling error in the eigenvalues is too high. An initial ensemble size of 576 was selected so that the variance of the sampling distribution for $\langle \hat{c}_k \rangle_N$ would not exceed 5% of its maximum value of 1 in the worst-case scenario where $\langle (\hat{c}_k)'^2 \rangle = 1$.

The relevant timescales depend on the frequency of the Fourier mode being evolved. Higher-frequency modes come to equilibrium with the forcing much more rapidly than those with lower-frequency. To account for this, the final time to which the high-fidelity model is evolved is scaled according to the wavelength of the Fourier mode used as the initial condition. The final time is defined to be the time required to advect a specified number

of wavelengths at the bulk velocity of 1. Thus for the Fourier mode $\exp(i2\pi k/L_x x)$ with wavelength L_x/k, the final time would be NL_x/k, where N is the number of wavelengths to travel. The number of snapshots in time collected for each eigenvalue is held constant as a function of k, so the interval between snapshots decreases with increasing k. For this work, 500 snapshots per advection of one wavelength were taken, and two wavelengths were traveled.

4.2 Generating a collection of scenarios

The goal is to develop a model for the dependence of $p(\widetilde{\boldsymbol{\lambda}})$ on the statistics of $\mathbf{u}$ over a range of scenarios that exhibit anomalous diffusion. The summary statistics $\langle u \rangle$, $\langle u'^2 \rangle$, $\langle v'^2 \rangle$, and ℓ (the integrated autocorrelation length of u, a measure of streamwise correlation) are postulated to be most important to anomalous diffusion. Correspondingly the nondimensional parameters $\langle u'^2 \rangle^{1/2}/\langle u \rangle$, $\langle v'^2 \rangle^{1/2}/\langle u \rangle$, and ℓ/L_x are chosen to define the scenario. The mean velocity $\langle u \rangle$ was chosen as a scale because its effect on the evolution of $\langle c \rangle$ is completely predictable. The integrated autocorrelation length ℓ was chosen because it is a characteristic length scale of the velocity.

To study the scenario dependence of $p(\widetilde{\boldsymbol{\lambda}})$, a representative collection of scenarios was defined. To this end, reasonable limits were placed on the space of nondimensional scenario parameters $\langle u'^2 \rangle^{1/2} / \langle u \rangle$, $\langle v'^2 \rangle / \langle u \rangle$, and ℓ/L_x. It is expected that for $\langle u'^2 \rangle^{1/2}/\langle u \rangle$ and $\langle v'^2 \rangle^{1/2}/\langle u \rangle$ smaller than 0.5, local fluctuations from the mean will be too weak to induce significant anomalous diffusion. Upper bounds of 1.5 and 1.0 were chosen for $\langle u'^2 \rangle^{1/2}/\langle u \rangle$ and $\langle v'^2 \rangle^{1/2}/\langle u \rangle$ respectively, based on empirical observations of the maximum values they achieved across a wide range of permeability statistics. A lower bound of 2% is placed on ℓ/L_x because for such short correlation lengths the separation of scales is large enough that a gradient-diffusion model is a good representation of the effects of dispersion. An upper bound of 10% is placed

on ℓ/L_x to ensure that the relevant statistical length-scales of the velocity are not so large that the assumption of periodicity is violated.

To study how $p(\widetilde{\boldsymbol{\lambda}})$ behaved as two scenario parameters were fixed and the third varied, a collection of ensembles spanning the scenario space in a coarse (4 points in each direction) grid was created. The mapping between the permeability covariance parameters ℓ_x, ℓ_y, and σ^2 and scenario parameters $\langle u'^2\rangle^{1/2}/\langle u\rangle$, $\langle v'^2\rangle^{1/2}/\langle u\rangle$ and ℓ/L_x is not known analytically, so a Gaussian process (GP) model was used to predict the permeability statistics that would produce the desired grid.

An initial training set for the GP was generated by defining a grid over reasonable ranges for the permeability parameters. The correlation lengths ℓ_x and ℓ_y were chosen to span from 1% to 10% of the domain length relative to the depthwise direction. The variance σ^2 affects the orders-of-magnitude variation in the sampled permeability field, so a range of $\sigma^2 \in [2, 13]$ was chosen to cover a wide range of cases in terms of variability. The GP was used to predict the permeability distribution parameters that would yield scenario parameters on the desired grid. Permeability ensembles were generated with the predicted parameters and used to create the corresponding velocity ensembles. The computed scenario parameters $\langle u'^2\rangle^{1/2}/\langle u\rangle$, $\langle v'^2\rangle^{1/2}/\langle u\rangle$, and ℓ/L_x for the velocity ensembles were then added to the training set of the GP. This process continued iteratively until ensembles with scenario parameters within 5% of the target grid points were found.

4.3 Analysis of ensemble statistics

Each scenario exhibits varying degrees of anomalous diffusion. To observe the difference in diffusion across scenarios, the mean evolution of a Gaussian initial condition was computed for each ensemble of velocities. The evolved means were compared by fixing two of the nondimensional scenario parameters and varying the third. As shown in Figure 4.1, varying

$\langle u'^2\rangle^{1/2}/\langle u\rangle$ yields significantly different evolutions, while varying $\langle v'^2\rangle^{1/2}/\langle u\rangle$ yields evolutions that are quite similar, indicating that it does not significantly impact the nature of the diffusion in the mean. Varying ℓ/L_x yields similar differences in the evolution to varying $\langle u'^2\rangle^{1/2}/\langle u\rangle$. Within the range of scenarios studied, an increase in either $\langle u'^2\rangle^{1/2}/\langle u\rangle$ or ℓ/L_x produced more anomalous diffusion, and an increase in both produced the greatest change, as shown in Figure 4.2. Based on these observations, analysis of the dependence of $p(\widetilde{\boldsymbol{\lambda}})$ on scenario parameters focused on $\langle u'^2\rangle^{1/2}/\langle u\rangle$ and ℓ/L_x.

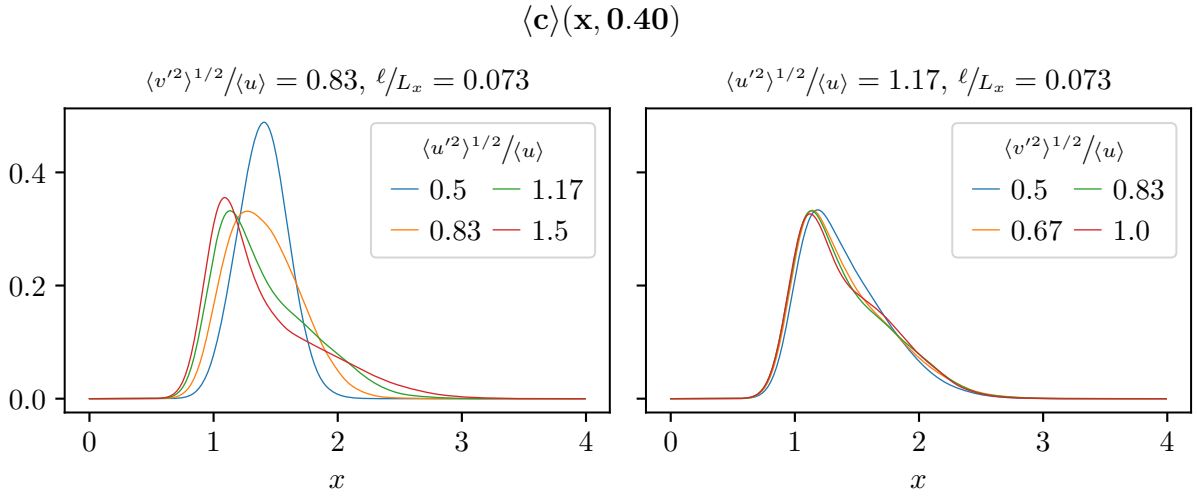

Figure 4.1: The mean concentration $\langle c\rangle$ after a Gaussian pulse is evolved to time $t = 0.40$, or 10% of a flowthrough time ($L_x/\langle u\rangle$).

The ultimate goal is to construct a relatively simple stochastic representation of $\widetilde{\mathcal{L}}$ as a function of scenario by modeling $p(\widetilde{\boldsymbol{\lambda}})$. Although the eigenvalues $\widetilde{\boldsymbol{\lambda}}$ are known to be functions of time, all but the eigenvalues associated with the largest length scales rapidly become stationary. Samples of eigenvalue evolutions for the two largest length scales are shown in Figure 4.3. Though some sample $\widetilde{\lambda}_k^{(i)}$ continue to vary with time, a majority become stationary. Note that while $\widetilde{\lambda}_1$ samples come to stationarity after approximately one flowthrough time (the time required to advect a domain length at the mean velocity, $L_x/\langle u\rangle$), $\widetilde{\lambda}_2$ samples require only 1/2 of the flowthrough time. This trend continues, with $\widetilde{\lambda}_3$ samples

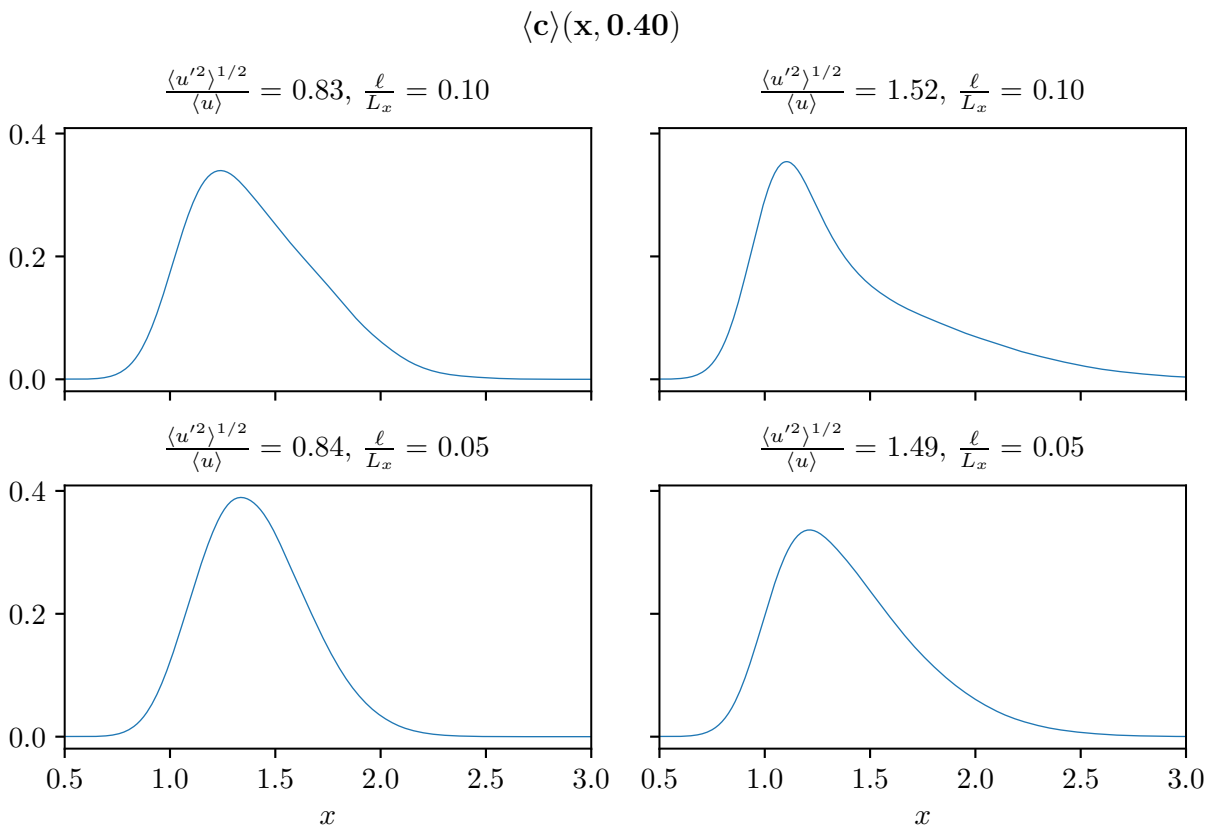

Figure 4.2: The mean concentration $\langle c \rangle$ after a Gaussian pulse is evolved to time $t = 0.40$, or 10% of a flowthrough time, for four different scenarios in terms of $\langle u'^2 \rangle^{1/2}/\langle u \rangle$ and ℓ/L_x.

coming to stationarity within 1/3 of the flowthrough time, $\widetilde{\lambda}_4^{(i)}$ within 1/4 of a flowthrough time, and so on. Based on this observation, the remainder of the analysis focuses on the stationary values of the eigenvalues. For simplicity of notation, for the remainder of this chapter $\widetilde{\boldsymbol{\lambda}}_{stationary} \equiv \widetilde{\boldsymbol{\lambda}}$. The remainder of this chapter examines key summary statistics of $p(\widetilde{\boldsymbol{\lambda}})$, namely its mean and covariance, and how they vary with $\langle u'^2 \rangle^{1/2}/\langle u \rangle$ and ℓ/L_x. Low-dimensional descriptions of these statistics and their dependence on $\langle u'^2 \rangle^{1/2}/\langle u \rangle$ and ℓ/L_x are sought.

Eigenvalue samples were computed for $k = 1$ through 20 to study their wavenumber dependence. Summary statistics of each ensemble, such as the sample mean and covariance of the eigenvalue samples, were computed. The sample mean of the eigenvalues represent

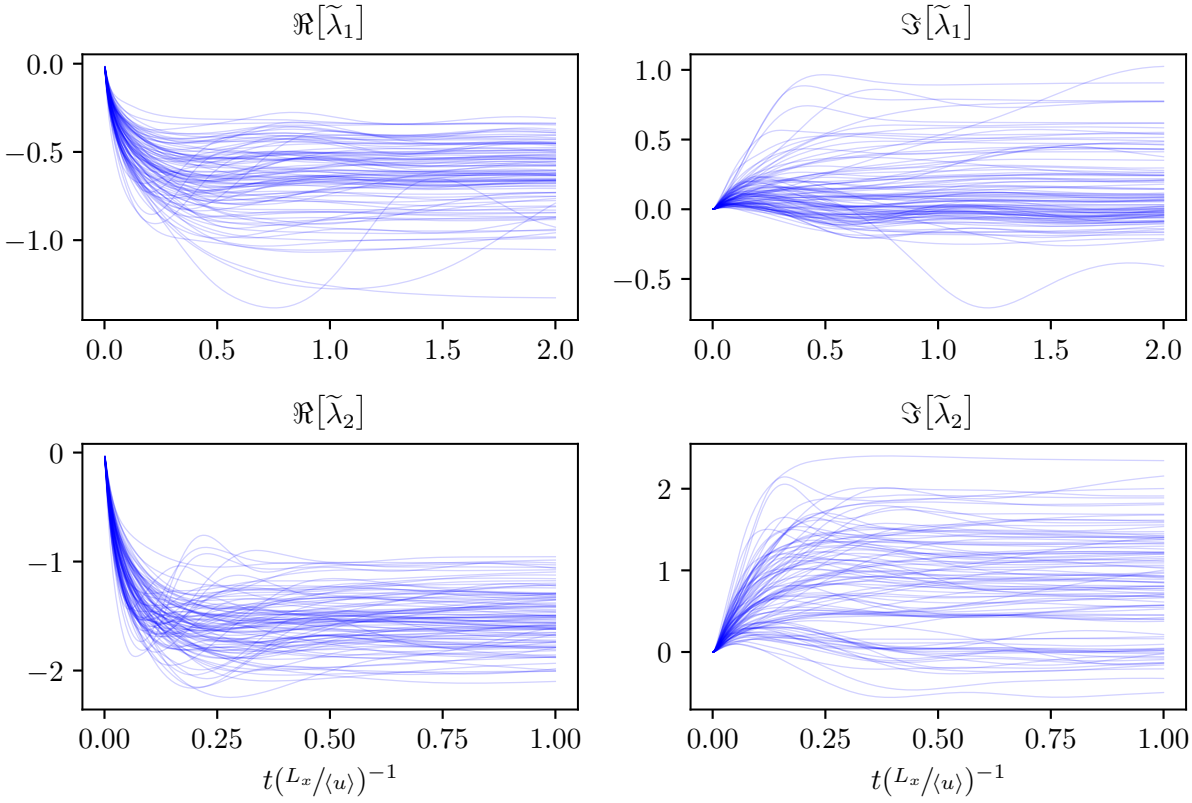

Figure 4.3: Samples $\widetilde{\lambda}_1^{(i)}$ and $\widetilde{\lambda}_2^{(i)}$ as a function of time, scaled by the flowthrough time.

the mean effect of dispersion on the state. Contrary to gradient-diffusion and fractional-derivative models of the dispersion, the eigenvalues did not grow as a fixed power of k (see Figure 4.4). Furthermore, the real and imaginary parts of $\mathbb{E}(\widetilde{\boldsymbol{\lambda}})$ do not exhibit the same dependence on k. For the cases exhibiting anomalous diffusion such as in Figure 4.4, the eigenvalues grew close to linearly as a function of k for higher wavenumbers, though the growth rates varied as a function of scenario.

The gradient-diffusion model of dispersion implies quadratic dependence on k and purely-real eigenvalues. As shown in Figure 4.5, scenarios with small $\langle u'^2 \rangle^{1/2} / \langle u \rangle$ and ℓ / L_x, where scale separation is more significant, the imaginary parts of the eigenvalues are negligible. The

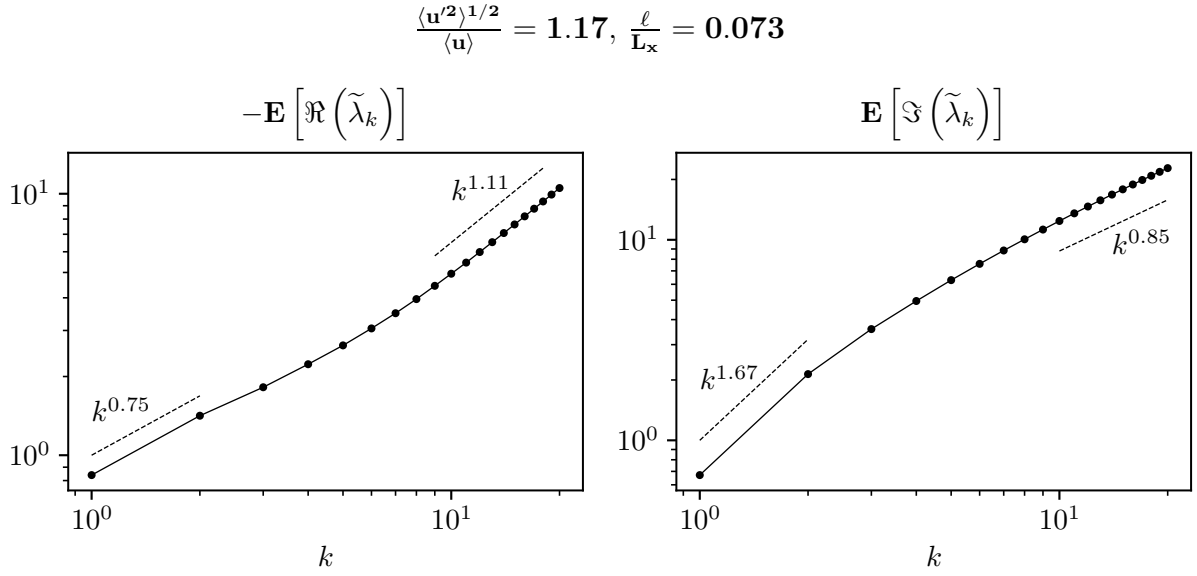

Figure 4.4: The k dependence of the real and imaginary parts of $\mathbb{E}\big(\Re\big[\widetilde{\lambda}_k\big]\big)$.

low-wavenumber growth rate for the real part of the eigenvalue grows as $\langle u'^2\rangle^{1/2}/\langle u\rangle$ and ℓ/L_x decrease, reaching $k^{1.78}$ for the lower bound of both scenario parameters $\left[\langle u'^2\rangle^{1/2}/\langle u\rangle, \ell/L_x\right] = [0.5, 0.02]$, but its dependence on k for higher wavenumbers deviates from the second derivative. This indicates that the diffusion is anomalous, but because it is only anomalous for higher wavenumbers, the effect is not apparent in the evolution of $\langle c\rangle$. For significantly anomalous cases, the imaginary parts of the eigenvalues are quite large, as shown in Figure 4.5. This indicates that dispersion has not only diffusive effects but significantly affects the advection of the mean state as well. This effect can be understood by rearranging the evolution equation of $\langle \hat{c}_k \rangle_y$:

$$\frac{\partial \langle \hat{c}_k \rangle_y}{\partial t} + i\left(\langle u\rangle (a_k) - \Im\big[\widetilde{\lambda}_k\big]\right)\langle \hat{c}_k \rangle_y = \left(-\nu_p a_k^2 + \Re\big[\widetilde{\lambda}_k\big]\right)\langle \hat{c}_k \rangle_y . \tag{4.8}$$

The advection velocity for each $\langle \hat{c}_k \rangle_y$ is $\langle u\rangle - \Im[\widetilde{\lambda}_k]/a_k$, so positive $\Im\big[\widetilde{\lambda}_k\big]$ result in advection that is slower than the bulk velocity. For anomalous cases, the advection velocity varies with k, as shown in Figure 4.6. For nonanomalous cases, the advection velocity is approximately

equal to $\langle u \rangle$.

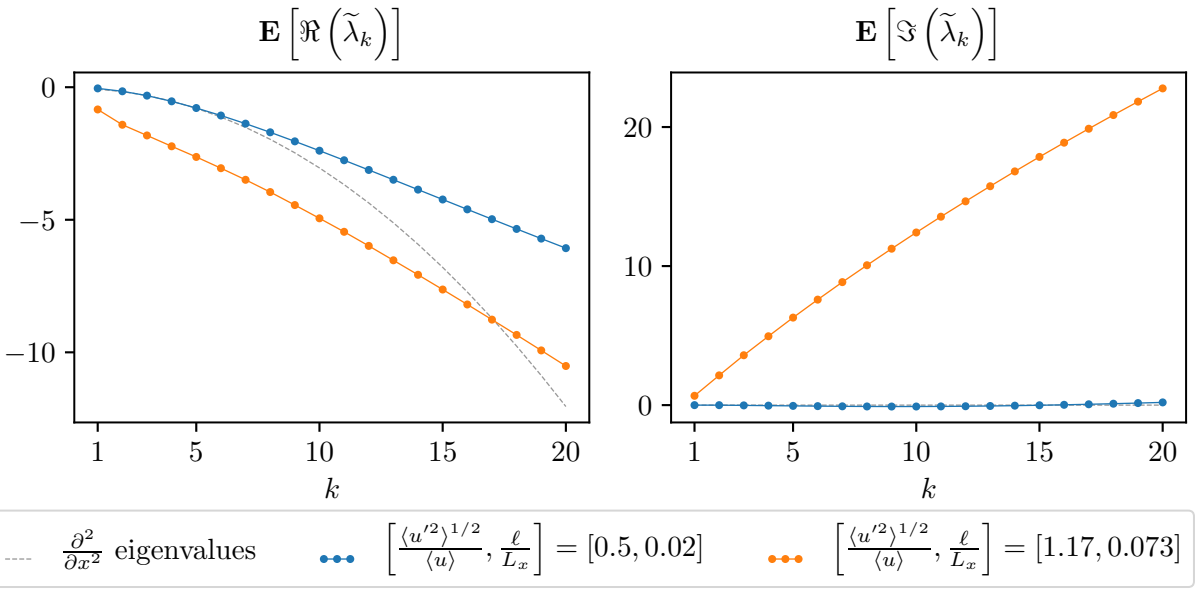

Figure 4.5: The real and imaginary parts of $\mathbb{E}(\widetilde{\boldsymbol{\lambda}})$ for two different scenarios defined in terms of $\langle u'^2\rangle^{1/2}/\langle u\rangle$ and ℓ/L_x.

The covariance of the eigenvalues across wavenumber as well as between real and imaginary parts was computed. To do this, the real and imaginary parts of the eigenvalue samples were included separately in sample vectors. Then the sample covariance is computed by

$$s_i \equiv \left[\Re[\widetilde{\lambda}_1^{(i)}], \Re[\widetilde{\lambda}_2^{(i)}], \cdots, \Im[\widetilde{\lambda}_1^{(i)}], \Im[\widetilde{\lambda}_2^{(i)}], \cdots\right] \in \mathbb{R}^{2N_k},$$

$$\Sigma_N \approx \frac{1}{N-1}\sum_{i=1}^{N}\left(s_i - \frac{1}{N}\sum_{j=1}^{N} s_j\right)^2,$$

where N_k is the number of eigenvalues and N is the number of samples. The resulting sample covariance is structured as follows:

$$\Sigma = \left[\begin{array}{c|c} \mathrm{Cov}\left(\Re[\widetilde{\boldsymbol{\lambda}}], \Re[\widetilde{\boldsymbol{\lambda}}]\right) & \mathrm{Cov}\left(\Re[\widetilde{\boldsymbol{\lambda}}], \Im[\widetilde{\boldsymbol{\lambda}}]\right) \\ \hline \mathrm{Cov}\left(\Im[\widetilde{\boldsymbol{\lambda}}], \Re[\widetilde{\boldsymbol{\lambda}}]\right) & \mathrm{Cov}\left(\Im[\widetilde{\boldsymbol{\lambda}}], \Im[\widetilde{\boldsymbol{\lambda}}]\right) \end{array}\right].$$

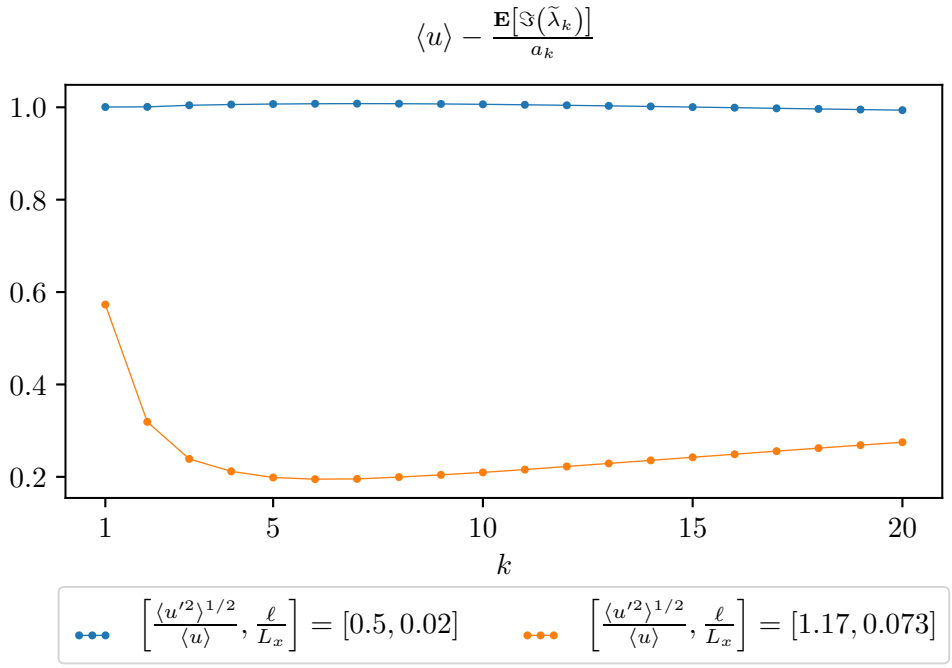

Figure 4.6: The advection velocity for two different scenarios defined in terms of $\langle u'^2\rangle^{1/2}/\langle u\rangle$ and ℓ/L_x.

Though some cases exhibited negative covariance between real and imaginary parts, the eigenvalues exhibited significant covariance across all scenarios, as shown in Figure 4.7. Negative covariance between real and imaginary parts was most pronounced for the smallest $\langle u'^2\rangle^{1/2}/\langle u\rangle$ scenarios, decreasing until it became positive for $\langle u'^2\rangle^{1/2}/\langle u\rangle \geq 1$. A low-rank approximation of the covariance was sought to minimize the number of degrees of freedom needed to model $p(\widetilde{\boldsymbol{\lambda}})$ in Chapter 5. The eigendecomposition of the covariance matrix for each scenario was computed, denoted $\Sigma = VWV^T$. It was found that the first two eigenvalues were dominant across all scenarios (see, e.g. Figure 4.8). This indicates the covariance admits a low-rank approximation using its first two eigenvalues and eigenvectors.

Additionally it was observed that consecutive eigenvector pairs $\{\mathbf{v_1}, \mathbf{v_2}\}$, $\{\mathbf{v_3}, \mathbf{v_4}\}$, etc. share similar information. This is best seen by considering the real and imaginary components of the i^{th} eigenvector of the covariance $\mathbf{v_i}$. Because the covariance was computed for an

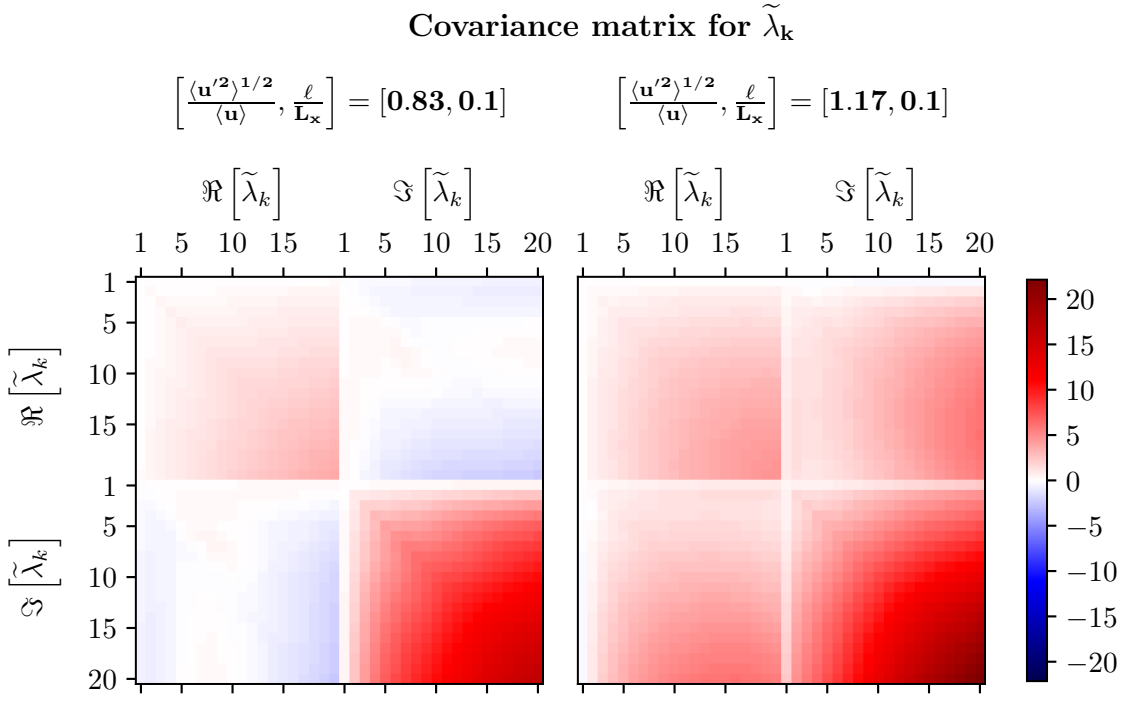

Figure 4.7: Covariance matrices of $p(\widetilde{\boldsymbol{\lambda}})$ for two scenarios with varying $\langle u'^2\rangle^{1/2}/\langle u\rangle$.

unraveled vector version of the complex eigenvalues, $\left[\Re[\widetilde{\boldsymbol{\lambda}}], \Im[\widetilde{\boldsymbol{\lambda}}]\right]$, the first half of $\mathbf{v_i}$'s components correspond to the real parts, the second half to the imaginary parts of $\widetilde{\boldsymbol{\lambda}}$. Each eigenvector can be accordingly split into real and imaginary contributions as $\mathbf{v}_i = [\mathbf{v_R^i}, \mathbf{v_I^i}]$. Written in complex form it was found that each eigenvector pair coarsely approximates the relation $\mathbf{v_1} \approx i\mathbf{v_2}$ (see, e.g. Figure 4.9). Assuming the equality $\mathbf{v_1} = i\mathbf{v_2}$, the first two eigenvectors and eigenvalues could be considered one complex eigenvalue, eigenvector pair as $\{w_1 + iw_2, \mathbf{v_R^1} + i\mathbf{v_I^1}\}$. The approximation is loose, but a rudimentary sensitivity analysis described in Section 5.1 found that it did not significantly impact the predicted evolution of $\langle c\rangle$. This fact is exploited to further reduce the rank of the covariance representation to $\mathbf{v_1}$ in Chapter 5.

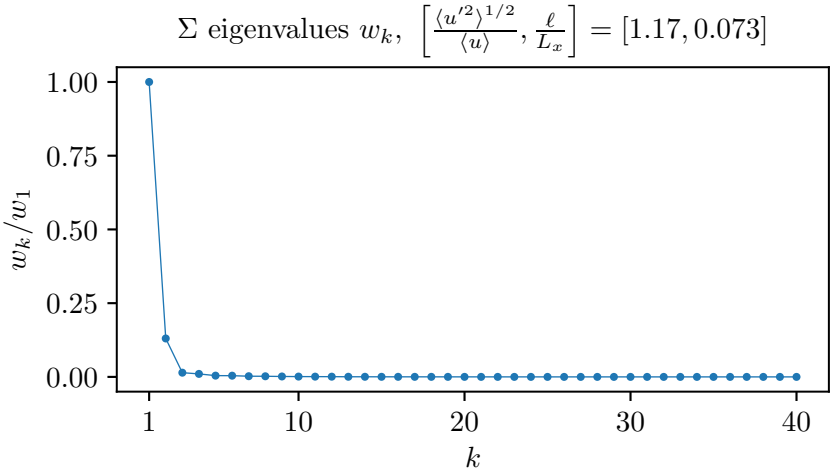

Figure 4.8: Eigenvalues of the covariance matrix of $p(\widetilde{\boldsymbol{\lambda}})$, Σ.

4.4 Conclusion

In this chapter, direct numerical simulation of eigenfunctions of the mean system through the detailed system enabled a more intrusive look at the dynamics of the mean system. By simulating the individual eigenfunctions rather than, e.g. a Gaussian pulse, it was possible to directly extract the corresponding eigenvalues of the operator representing dispersion. The computation of ensembles of eigenvalues over a range of scenarios in terms of velocity statistics allowed for the study of their distribution $p(\widetilde{\boldsymbol{\lambda}})$ and how it varied with scenario. Key statistics of $p(\widetilde{\boldsymbol{\lambda}})$ such as mean and covariance appear to vary predictably with $\langle u'^2\rangle^{1/2}/\langle u\rangle$ and ℓ/L_x, and the low rank of the covariance of the eigenvalues suggests that $p(\widetilde{\boldsymbol{\lambda}})$ may admit a relatively low-dimensional representation. In Chapter 5, the feasibility of such a low-dimensional representation depending on only these two scenario parameters will be explored.

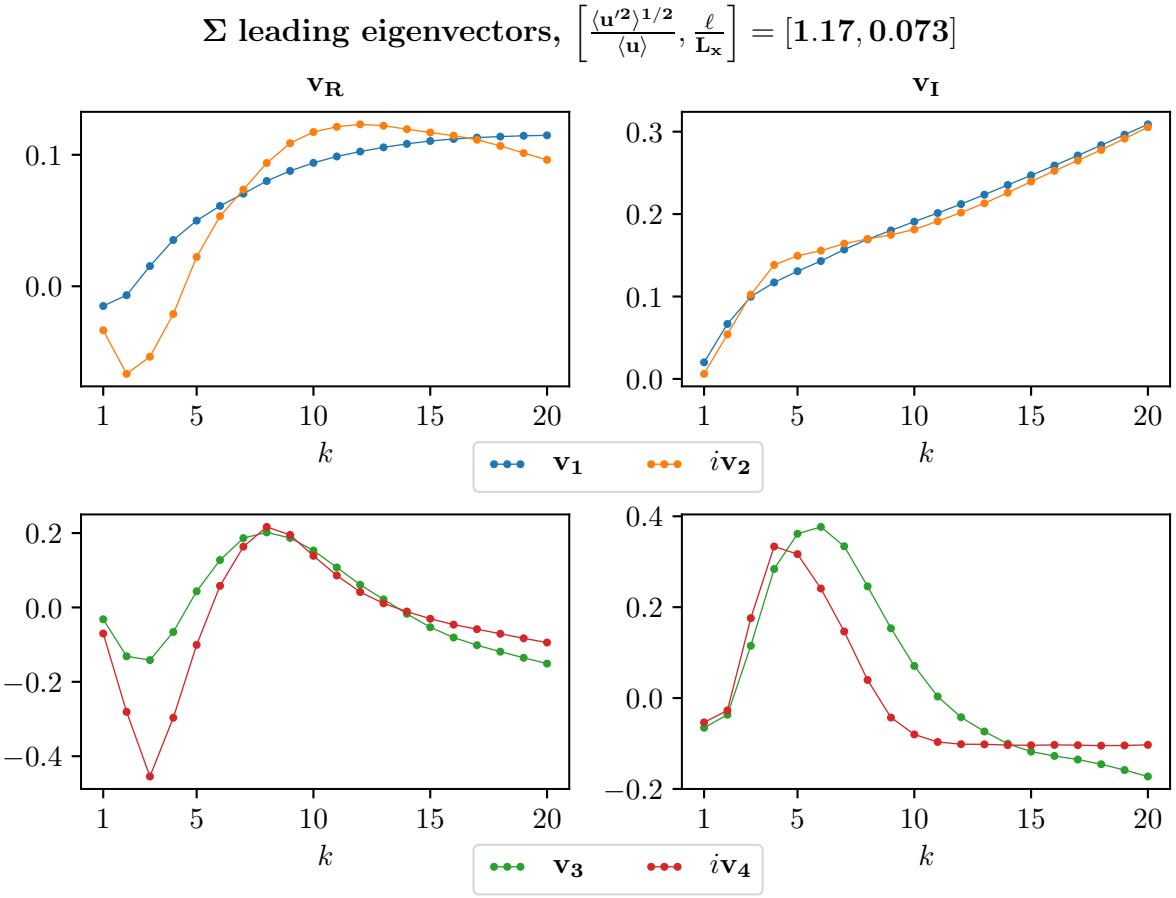

Figure 4.9: Complex forms of the eigenvectors of the covariance matrix of $p(\widetilde{\boldsymbol{\lambda}})$, Σ, plotted in pairs $\{\mathbf{v_1}, i\mathbf{v_2}\}$ and $\{\mathbf{v_3}, i\mathbf{v_4}\}$.

Chapter 5

Stochastic formulation of model-form uncertainty

In this chapter, the deterministic formulation derived in Chapter 2 and the results of the study in Chapter 4 are combined to define a model-form uncertainty representation for missing microstructural information in the mean evolution equation of $\langle c \rangle$. The modeling requirements placed on the representation are that it should depend on the state variable $\langle c \rangle$, should be able to extrapolate with respect to scenario, that it should be stochastic, and that it should not violate the physics of the problem. Each of these constraints will be considered in the development of the model-form uncertainty representation.

The representation is defined in terms of $\widetilde{\mathcal{L}}$, which acts on individual samples of the depthwise-averaged state $\langle c \rangle_y$. The action of $\widetilde{\mathcal{L}}$ on $\langle c \rangle_y$ is defined in terms of its eigenvalues $\widetilde{\boldsymbol{\lambda}}$ and eigenfunctions, the Fourier modes. Uncertainty in the model is expressed through the probability distribution of its eigenvalues, $p\big(\widetilde{\boldsymbol{\lambda}}\big)$. To impose stochasticity on the distribution, it is specified in terms of hyperparameters $\boldsymbol{\xi}$ and is denoted $p\big(\widetilde{\boldsymbol{\lambda}}; \boldsymbol{\xi}\big)$. Information from data can be used to update $\boldsymbol{\xi}$, but not the mean and covariance of $p\big(\widetilde{\boldsymbol{\lambda}}\big)$ directly. This guarantees that in the limit of infinite data, the result of calibration would be an optimal distribution for $\widetilde{\boldsymbol{\lambda}}$, not a Dirac delta around optimal values of $\widetilde{\boldsymbol{\lambda}}$. Uncertainty in the values of $\boldsymbol{\xi}$ motivates casting them as random variables and posing distributions for them.

Forward propagation of uncertainty to the evolution of the mean concentration $\langle c \rangle = \mathbb{E}\left[\langle c \rangle_y\right]$ proceeds as follows. Sample hyperparameters $\boldsymbol{\xi}^{(i)}$ are drawn from their distribu-

tions and are used to specify $p(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi})$, from which sample eigenvalues $\widetilde{\boldsymbol{\lambda}}^{(i)}$ are drawn. The corresponding evolution of $\langle c\rangle_y^{(i)}$ is computed cheaply using its Fourier series solution,

$$\langle c\rangle_y^{(i)}(x,t) = 2\Re\left[\sum_{k=0}^{N_x/2}\langle\hat{c}_k\rangle_y(0)\exp\left(\left[-\nu_p(a_k)^2+\widetilde{\lambda}_k^{(i)}+\langle u\rangle\right]t+(ia_k)x\right)\right],$$

where $\langle\hat{c}_k\rangle_y(0)$ is the k^{th} Fourier coefficient of the initial condition and $a_k = {}^{2\pi k}/_{L_x}$. The sample evolutions $\langle c\rangle_y^{(i)}$ are then used to compute the sample mean and variance of the predicted evolution of the mean $\langle c\rangle$.

First, an appropriate hyperparametrization of $p(\widetilde{\boldsymbol{\lambda}})$ must be specified. In Section 5.1, representations of $p(\widetilde{\boldsymbol{\lambda}})$ with different simplifying assumptions are used for forward propagation to determine which aspects of $p(\widetilde{\boldsymbol{\lambda}})$ are most important for prediction. These findings are used in Section 5.2 to define the stochastic representation of $p(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi})$ in terms of its hyperparameters $\boldsymbol{\xi}$, and scenario dependence is imposed on $\boldsymbol{\xi}$. Finally, in Section 5.3 the results of forward propagation with the model-form uncertainty representation are presented and discussed.

5.1 Sensitivity analysis

As discussed in Chapter 4, the distribution of the eigenvalues, $p(\widetilde{\boldsymbol{\lambda}})$, appears to admit a low-dimensional representation. A rudimentary sensitivity analysis was performed to determine which aspects of the distribution were most important in determining the mean concentration's evolution. This was done by posing several approximate representations of $p(\widetilde{\boldsymbol{\lambda}})$, each with their own simplifying assumption, then performing forward propagation to $\langle c\rangle$ for each one. For instance, one approximate distribution assumed independence between the eigenvalues and was fit to reproduce the marginal distributions of each eigenvalue individually. Another used a low-rank approximation of the covariance of $p(\widetilde{\boldsymbol{\lambda}})$. If an approximation of

$p(\widetilde{\boldsymbol{\lambda}})$ resulted in nonphysical samples of $\langle c \rangle_y$, or resulted in predictions of $\langle c \rangle$ that were in poor agreement with the $\langle c \rangle$ computed from the high-fidelity model, it was deemed invalid, and the aspect of $p(\widetilde{\boldsymbol{\lambda}})$ that was modeled away was deemed important for determining $\langle c \rangle$.

Before defining approximate representations of $p(\widetilde{\boldsymbol{\lambda}})$, the stationary values of the eigenvalues in each ensemble were used for forward propagation to $\langle c \rangle$. Since they are distributed exactly according to $p(\widetilde{\boldsymbol{\lambda}})$ without any approximations besides the assumption of being constant in time, the effect of not accounting for time dependence can be assessed. In Figure 5.1, the sample mean for $\langle c \rangle$ and 95% confidence interval generated using the sample evolutions from the high-fidelity model are compared to the estimated evolutions using the stationary values of $\widetilde{\boldsymbol{\lambda}}$. The 95% confidence interval is estimated using a normality assumption for the distribution of $\langle c \rangle_y$, with its standard deviation σ computed as the square root of the sample variance. The normality assumption is inexact, as evidenced by the 95% confidence interval for the DNS data assigning significant probability to negative concentrations, though no sample evolutions from the high-fidelity model yielded negative concentrations. A more sophisticated estimate of the 95% confidence interval would be necessary to ameliorate this nonphysical estimate of the distribution of $\langle c \rangle_y$. However, the normal approximation is sufficient as a rough approximation of the amount of variation in $\langle c \rangle$.

As shown in Figure 5.1, although the mean evolution is captured well, there is more variance in the DNS data than from propagating the stationary values of $\widetilde{\boldsymbol{\lambda}}$. This must be due to neglecting the transient evolution of the eigenvalues before they become stationary, since the sample eigenvalues are otherwise exact. To account for this, the time history of the eigenvalues or of the evolution of the variance $\langle c'^2 \rangle$ would also need to be modeled along with their own uncertainty representation. Such developments are left to future work as refinements of the current formulation.

To assess the importance of accounting for the covariance of the eigenvalues, a distribution for the eigenvalues was posed that assumed independence across wavenumber and between

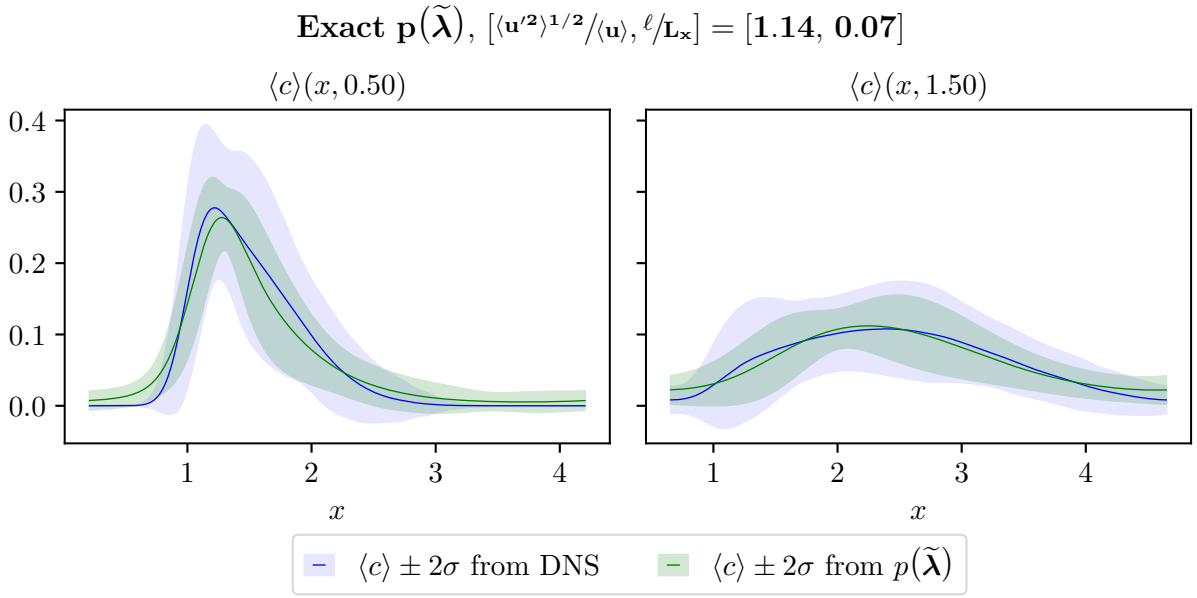

Figure 5.1: The mean and 95% confidence interval of $\langle c \rangle$ computed by direct numerical simulation (DNS), compared to the mean and 95% confidence interval of $\langle c \rangle$ predicted from computed stationary values of the eigenvalues $\widetilde{\boldsymbol{\lambda}}$.

real and imaginary parts of the eigenvalues. For the first 20 eigenvalues whose samples were computed directly, a skew normal distribution was fit to each $\Re[\widetilde{\lambda}_k]$ and $\Im[\widetilde{\lambda}_k]$, reproducing the shape of their marginal distributions. The skew normal distribution is a generalization of the normal distribution that allows for non-zero skewness, first introduced in [36]. Its distribution is defined as the product of a normal probability density function with a normal cumulative distribution function. It is specified in terms of a location parameter μ, a scale parameter ω, and a shape parameter α as

$$p(x) = 2\left[\frac{1}{\sqrt{2\pi\omega^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\omega^2}\right)\right]\left[\int_{-\infty}^{\alpha\frac{x-\mu}{\omega}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t\right].$$

The shape parameter controls the skewness of the distribution, and for $\alpha = 0$, the skew normal distribution simplifies to the normal distribution. In this limit μ and ω converge to the mean and variance of a normal distribution, and they play similar roles in the distribution

for $\alpha \neq 0$. For the sensitivity analysis, μ, ω and α were fit to the computed eigenvalues using a maximum likelihood optimization implemented in SciPy. The directly-computed eigenvalues' samples are compared to the fitted distribution in Figure 5.2.

Higher wavenumber eigenvalues with no corresponding computed values were modeled as independent Gaussians. Their means and variances as a function of k were approximated using linear models fit to the first 20 eigenvalues for which samples were computed directly. As shown in Figure 5.3, samples from this distribution produced $\langle c \rangle_y$ with non-physical oscillations in their tails and significant negative values in the concentration field. A distribution imposing correlation between the real and imaginary parts of eigenvalues of the same wavenumber but assuming independence across wavenumbers produced similar oscillations and negative concentrations. Based on these findings it was determined that accounting for covariance across wavenumbers was essential to producing physical sample evolutions of $\langle c \rangle_y$.

To account for the covariance of the eigenvalues across wavenumber and between real and imaginary parts, an approximation of $p(\widetilde{\boldsymbol{\lambda}})$ was posed as a multivariate normal distribution. The goal was to determine the simplest possible model for the behavior of its mean and covariance matrix. In Chapter 4 it was observed that $\mathbb{E}\big(\Re[\widetilde{\lambda}_k]\big)$ and $\mathbb{E}\big(\Im[\widetilde{\lambda}_k]\big)$ grew nearly linearly with k for scenarios exhibiting anomalous diffusion. As a result a simple linear model in k was posed to approximate their means. As was also noted in Chapter 4, across all scenarios the covariance matrix admitted a low-rank eigendecomposition approximation using only its first two eigenvalue-eigenvector pairs. Furthermore, the second eigenvector was approximately equal to the first eigenvector multiplied by the imaginary unit when written in a complex form. That is, $\mathbf{v^2} \approx i\,(\mathbf{v^1_R} + i\mathbf{v^1_I}) = -\mathbf{v^1_I} + i\mathbf{v^1_R}$. In an unraveled form, $\mathbf{v_2} = [\mathbf{v^2_R}\ \mathbf{v^2_I}] \approx [-\mathbf{v^1_I}\ \mathbf{v^1_R}]$. This approximation was used to define $\mathbf{v_2}$ in terms of $\mathbf{v_1}$. The real and imaginary parts of $\mathbf{v_1}$ were also each modeled as linear in k. The summary statistics of the first 20 eigenvalues for which samples were computed directly were used to compute

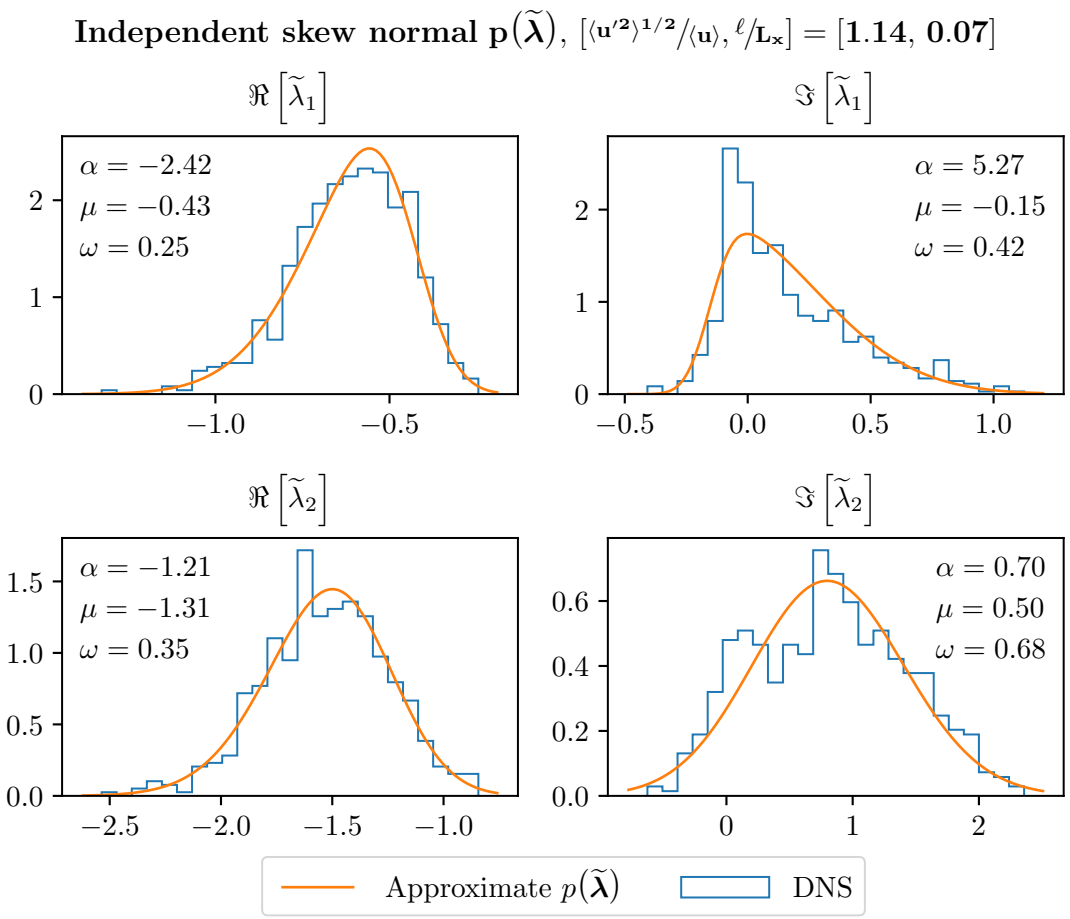

Figure 5.2: Independent skew normal distributions fit to directly computed samples of $\widetilde{\boldsymbol{\lambda}}$, plotted with histograms of the DNS samples.

the slopes and intercepts of the linear models. The computed eigenvalues of the covariance w_1 and w_2 were used directly. The multivariate-Gaussian linear approximation of $p(\widetilde{\boldsymbol{\lambda}})$ is thus defined as

$$\begin{aligned} p(\widetilde{\boldsymbol{\lambda}}) &= \mathcal{N}\Big(\mathbf{m}_{\boldsymbol{\lambda_R}} + i\mathbf{m}_{\boldsymbol{\lambda_I}}, \Sigma_2\Big), \\ \Sigma_2 &= w_1 \mathbf{v_1} \mathbf{v_1}^T + w_2 \mathbf{v_2} \mathbf{v_2}^T \\ \mathbf{v_1} &= [\mathbf{m}_{\mathbf{v_R}}\ \mathbf{m}_{\mathbf{v_I}}], \quad \mathbf{v_2} = [-\mathbf{m}_{\mathbf{v_I}}\ \mathbf{m}_{\mathbf{v_R}}], \\ (\mathbf{m}_\alpha)_k &= a_\alpha(k-1) + b_\alpha, \quad k \in [1, N_k], \quad \alpha \in \{\boldsymbol{\lambda_R}, \boldsymbol{\lambda_I}, \mathbf{v_R}, \mathbf{v_I}\}. \end{aligned} \tag{5.1}$$

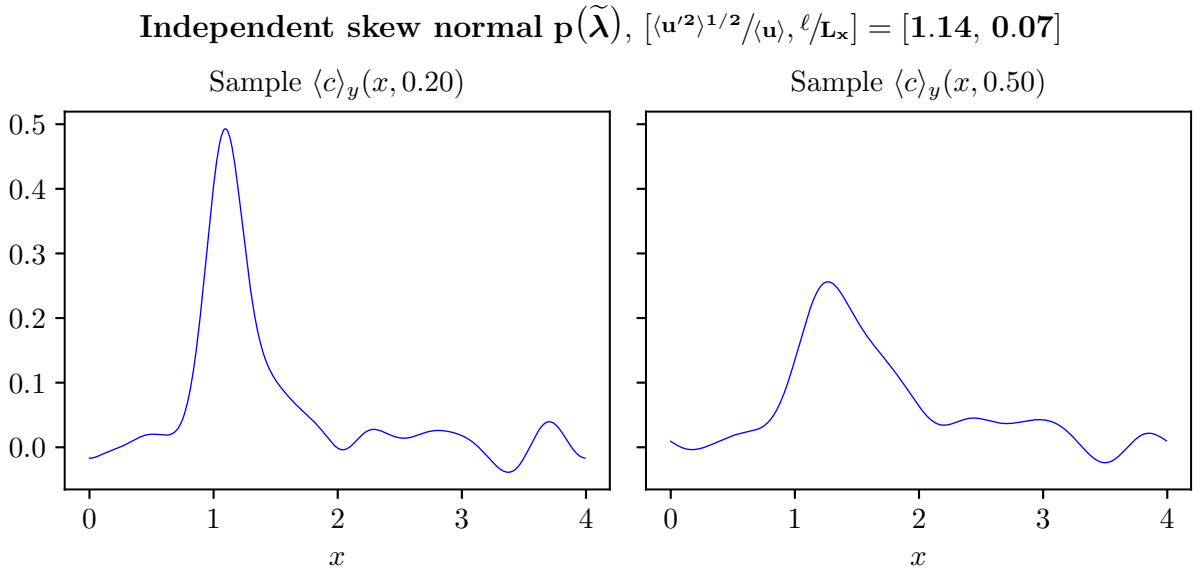

Figure 5.3: The depthwise-averaged state $\langle c \rangle_y$ evolved with $\widetilde{\boldsymbol{\lambda}}$ sampled from independent skew normal distributions fit to reproduce the marginal distributions of the exact $p(\widetilde{\boldsymbol{\lambda}})$.

Remarkably, this simple representation yields a reasonable approximation of $\langle c \rangle$ for the scenarios that exhibit anomalous diffusion (see, e.g. Figure 5.4). The mean is included in the high-probability region of the push-forward of $p(\widetilde{\boldsymbol{\lambda}})$ across most of the domain, although parts of the tails are not captured fully at time $t = 0.5$. Furthermore, accounting for covariance mitigates the nonphysical oscillations in the tails of $\langle c \rangle_y$, as shown in Figure 5.5. Though $\langle c \rangle_y$ is not prohibited from negative values by accounting for covariance, any negative values are much smaller than for any cases where the covariance was ignored.

For the scenarios with smaller values of $\langle u'^2 \rangle^{1/2} / \langle u \rangle$ and ℓ / L_x, which do not exhibit anomalous diffusion, the linear models of the mean real and imaginary parts of the eigenvalues are not as good approximations. This leads to poorer predictions of the evolution of $\langle c \rangle$, as shown in Figure 5.6. Although the approximation accurately predicts a Fickian diffusion of $\langle c \rangle$, the linear model for $\Re[\widetilde{\boldsymbol{\lambda}}]$ predicts larger negative contributions than the true values, which induces more rapid decay in $\langle c \rangle_y$. Given its success across a range of anomalous scenarios, however, this low-dimensional representation of $p(\widetilde{\boldsymbol{\lambda}})$ was used for the stochastic

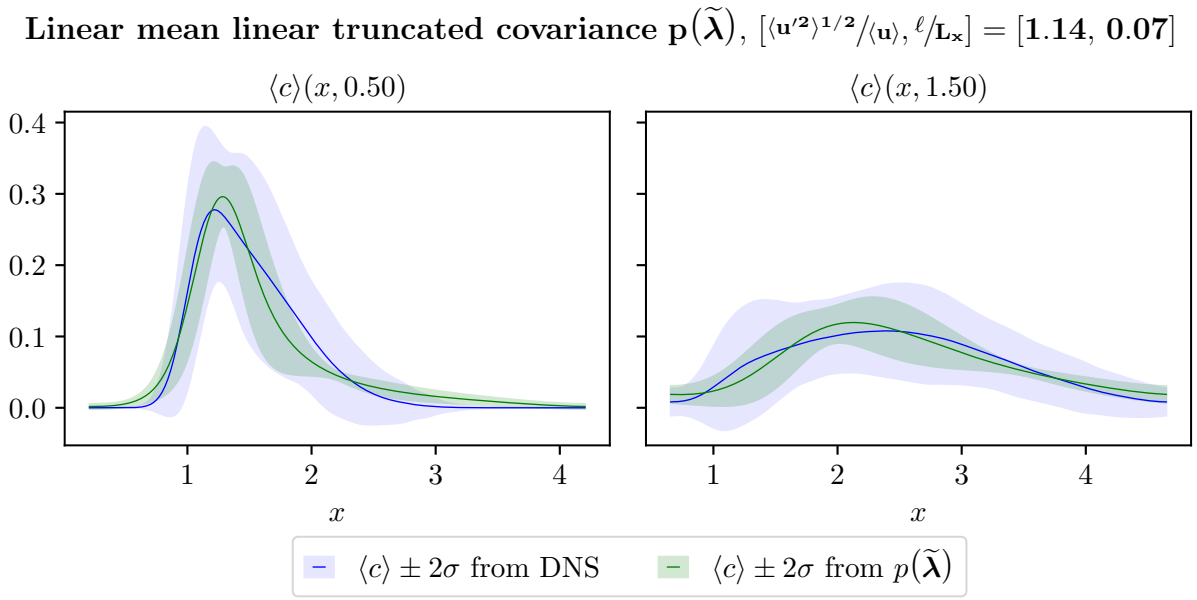

Figure 5.4: The mean and 95% confidence interval of $\langle c \rangle$ computed by DNS, compared to the mean and 95% confidence interval predicted by the linear, low-rank approximation of $p(\widetilde{\boldsymbol{\lambda}})$ defined in (5.1).

formulation of $\widetilde{\mathcal{L}}$.

5.2 Stochastic specification

To this point, the assumptions made in the formulation of $p(\widetilde{\boldsymbol{\lambda}})$ include: the eigenvalues $\widetilde{\boldsymbol{\lambda}}$ do not vary with time; the real and imaginary parts of their means vary linearly with k; there is no dependence on $\langle v'^2 \rangle^{1/2} / \langle u \rangle$; the $\widetilde{\boldsymbol{\lambda}}$ covariance matrix is described by its first two eigenvectors, the second being expressed in terms of the first. The first eigenvector of the covariance, $\mathbf{v_1}$, is split into the contributions from the real and imaginary parts as $\mathbf{v_1} = [\mathbf{v_R^1}, \mathbf{v_I^1}]$, and defined in (5.1). As was done in (5.1), the eigenvector's components are assumed to grown linearly in terms of their index k.

With these assumptions the low-dimensional representation of $p(\widetilde{\boldsymbol{\lambda}})$ defined in (5.1) is used as the basis of the stochastic formulation of $\widetilde{\mathcal{L}}$. This reduced representation depends on only 10 hyperparameters, collectively denoted $\boldsymbol{\xi}$ for brevity: $\{a_{\boldsymbol{\lambda_R}}, b_{\boldsymbol{\lambda_R}}\}$, the slope and

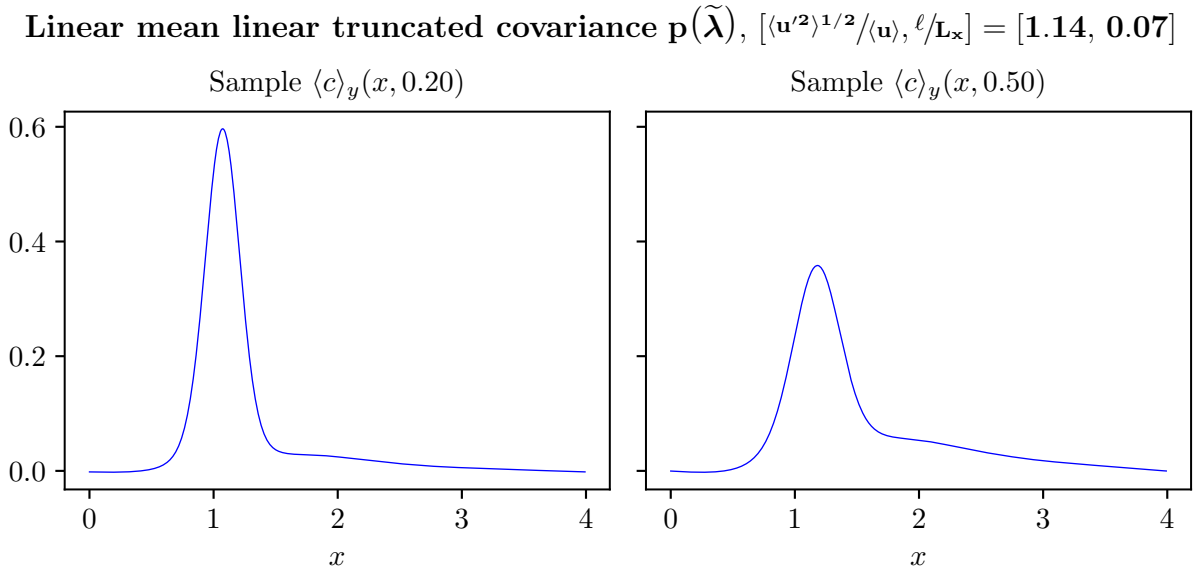

Figure 5.5: The depthwise-averaged state $\langle c \rangle_y$ evolved with $\widetilde{\boldsymbol{\lambda}}$, sampled from the linear, low-rank approximation of $p(\widetilde{\boldsymbol{\lambda}})$ defined in (5.1), for an anomalous scenario.

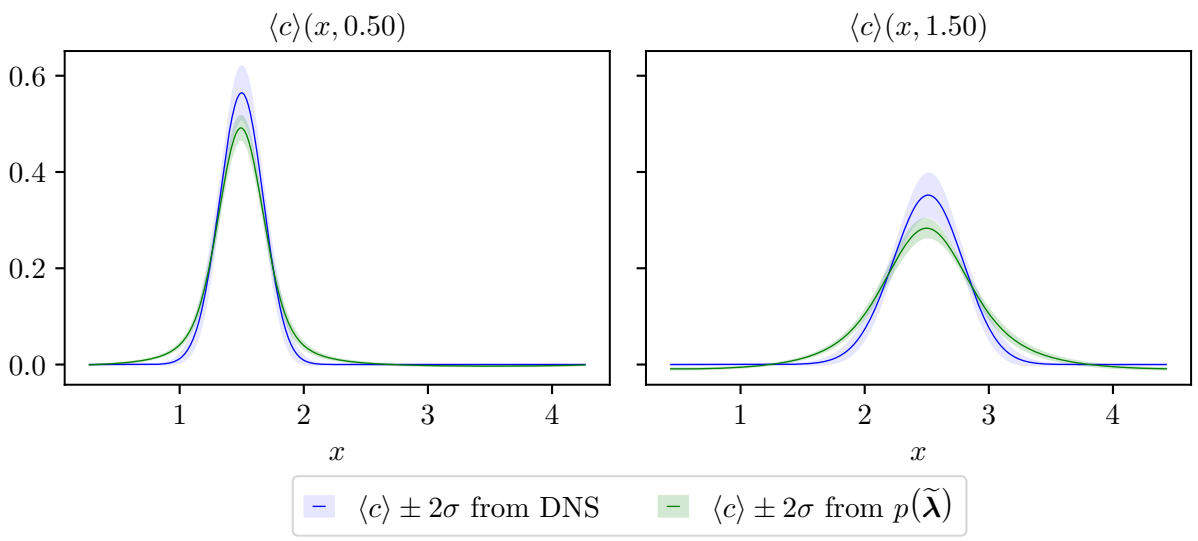

Figure 5.6: The mean and 95% confidence interval of $\langle c \rangle$ computed by DNS, compared to the mean and 95% confidence interval predicted by the linear, low-rank approximation of $p(\widetilde{\boldsymbol{\lambda}})$ defined in (5.1), for a nonanomalous scenario.

intercept of $\Re\big[\widetilde{\boldsymbol{\lambda}}\big]$; $\{a_{\boldsymbol{\lambda_I}}, b_{\boldsymbol{\lambda_I}}\}$, the slope and intercept of $\Im\big[\widetilde{\boldsymbol{\lambda}}\big]$; $\{a_{\mathbf{v_R}}, b_{\mathbf{v_R}}\}$, the slope and intercept of the real part of the first eigenvector of the covariance matrix, $\mathbf{v_R^1}$; $\{a_{\mathbf{v_I}}, b_{\mathbf{v_I}}\}$, the slope and intercept of the imaginary part of the first eigenvector of the covariance matrix, $\mathbf{v_I^1}$; and $\{w_1, {}^{w_1}\!/_{w_2}\}$, the first eigenvalue of the covariance matrix and its ratio with the second eigenvalue.

The deterministic constraints on $\mathcal{L}$ and subsequently on $\widetilde{\mathcal{L}}$ discussed in Chapter 2 will be taken into account in the statistical treatment of the hyperparameters. To review, the constraints to be satisfied are: $\widetilde{\lambda}_0 = 0$; $\Re\big[\widetilde{\lambda}_k\big] \leq 0$ for $k \geq 1$; and preservation of positivity. Note that the original constraint for enforcing decay in the evolution of $\langle c \rangle_y$ was $-\nu_p a_k^2 + \Re\big[\widetilde{\lambda}_k\big] \leq 0$, but that $\Re\big[\widetilde{\lambda}_k\big] \leq 0$ is a sufficient condition. The first constraint is satisfied trivially by setting $\widetilde{\lambda}_0$ to 0 deterministically. The constraint of positivity is not guaranteed to be satisfied by accounting for covariance in the eigenvalues, but empirically it mitigates the worst of the strong oscillations that produce large negative concentrations. The requirement that $\Re\big[\widetilde{\lambda}_k\big] \leq 0$ for $k \geq 1$ must be encoded in the slope and intercept of its linear model.

Models for $\boldsymbol{\xi}$ as a function of ${\langle u'^2 \rangle^{1/2}}/{\langle u \rangle}$ and ${\ell}/{L_x}$ were developed to encode scenario dependence in $p\big(\widetilde{\boldsymbol{\lambda}}; \boldsymbol{\xi}\big)$. However, ${\langle u'^2 \rangle^{1/2}}/{\langle u \rangle}$ and ${\ell}/{L_x}$ are only proxies for the underlying dependence of dispersion on local fluctuations of $\mathbf{u}$ from its mean. Because of this, scenario-based models of $\boldsymbol{\xi}$ on ${\langle u'^2 \rangle^{1/2}}/{\langle u \rangle}$ and ${\ell}/{L_x}$ will not perfectly capture this dependence, and uncertainty will remain. To account for this remaining uncertainty, each hyperparameter is considered uncertain and normally distributed. Their means and variances are defined to depend on ${\langle u'^2 \rangle^{1/2}}/{\langle u \rangle}$ and ${\ell}/{L_x}$.

To develop a representation of how the distribution of each hyperparameter depends on scenario, each hyperparameter is computed over the grid of ${\langle u'^2 \rangle^{1/2}}/{\langle u \rangle}$ and ${\ell}/{L_x}$ discussed in Chapter 4. The primary direction of variation in ${\langle u'^2 \rangle^{1/2}}/{\langle u \rangle}$ - ${\ell}/{L_x}$ space is identified for each hyperparameter, and the computed values are projected in that direction. A low-order (at most 3^{rd}-order) polynomial is then fit to the projected data and used as the mean of the

distribution. The variance is derived from the deviation of the computed values from this modeled mean. This process is detailed in Appendix E.

Prior information about how the eigenvalues $\widetilde{\boldsymbol{\lambda}}$ behave as a function of scenario was encoded in the form of the polynomial fits. For instance, it is known that as $\langle u'^2\rangle^{1/2}/\langle u\rangle$ and ℓ/L_x approach zero, the effect of dispersion on $\langle c\rangle$ vanishes. This implies that the mean and variance of $p\big(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi}\big)$ should vanish in the limit as well. To account for this, a zero intercept in terms of $\langle u'^2\rangle^{1/2}/\langle u\rangle$ and ℓ/L_x was enforced in the polynomial fits for $\{a_{\boldsymbol{\lambda_R}}, b_{\boldsymbol{\lambda_R}}, a_{\boldsymbol{\lambda_I}}, b_{\boldsymbol{\lambda_I}}, w_1\}$, the hyperparameters for the linear models of $\mathbb{E}\big(\widetilde{\boldsymbol{\lambda}}\big)$ and the first eigenvalue of the covariance matrix of $p\big(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi}\big)$. Furthermore, the uncertainty in their distributions should also decay to zero in the limit, since their values are known to be zero exactly. To account for this, the standard deviations of the distributions for $\{a_{\boldsymbol{\lambda_R}}, b_{\boldsymbol{\lambda_R}}, a_{\boldsymbol{\lambda_I}}, b_{\boldsymbol{\lambda_I}}, w_1\}$ approach zero with $\langle u'^2\rangle^{1/2}/\langle u\rangle$ and ℓ/L_x as well (see Appendix E for more details).

The constraint $\Re\big[\widetilde{\lambda}_k\big] \leq 0$ for $k \geq 1$ is satisfied if and only if $a_{\boldsymbol{\lambda_R}}, b_{\boldsymbol{\lambda_R}} \leq 0$. To ensure that $\{a_{\boldsymbol{\lambda_R}}, b_{\boldsymbol{\lambda_R}}\}$ simultaneously satisfied this constraint as well as vanishing in the lower limit, $(-a_{\boldsymbol{\lambda_R}})^{1/2}$ and $(-b_{\boldsymbol{\lambda_R}})^{1/2}$ were instead assumed to be normally distributed. Similarly, since the eigenvalues of the covariance matrix are known to be nonnegative, $(w_1)^{1/2}$ and $(w_1/w_2)^{1/2}$ were assumed normally distributed. Polynomial fits to $\langle u'^2\rangle^{1/2}/\langle u\rangle$ and ℓ/L_x were computed for the means of the distributions for $(-a_{\boldsymbol{\lambda_R}})^{1/2}$, $(-b_{\boldsymbol{\lambda_R}})^{1/2}$, $(w_1)^{1/2}$ and $(w_1/w_2)^{1/2}$. For a full definition of the stochastic formulation of $p\big(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi}\big)$, see Appendix F.

5.3 Results

As shown in Figure 5.7, even this simple stochastic representation of the model-form uncertainty is able to reasonably predict the evolution of $\langle c\rangle$ for $\langle u'^2\rangle^{1/2}/\langle u\rangle$ and ℓ/L_x away from the extremes of the scenario grid chosen for the study. The 95% confidence interval of the eigenvalues with this representation encapsulates the mean computed from their ensemble

directly, as shown in Figure 5.8. The true evolution of $\langle c \rangle$ is not completely encapsulated in the 95% confidence region of of the push-forward of $p(\widetilde{\boldsymbol{\lambda}}; \boldsymbol{\xi})$ across the entire domain at all times, but its value at the downstream boundary of the domain is encapsulated, as shown in Figure 5.9. For the purposes of predicting the time at which a contaminant would exceed a safe threshold downstream, this would be the most important quantity of interest.

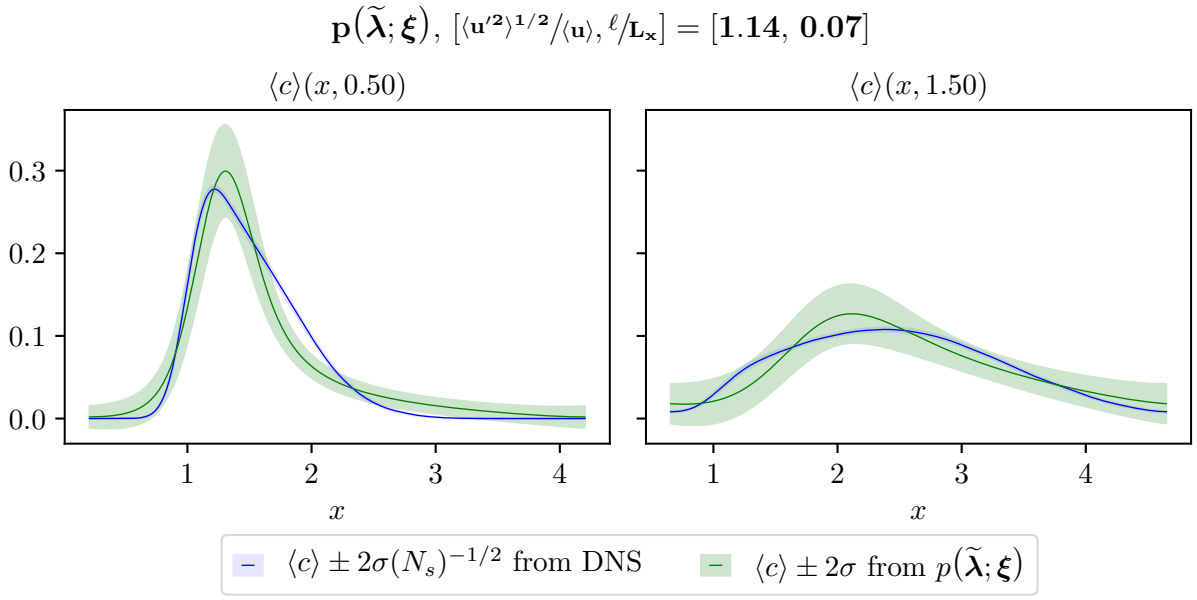

Figure 5.7: The push-forward of the stochastic formulation of $p(\widetilde{\boldsymbol{\lambda}}; \boldsymbol{\xi})$ to $\langle c \rangle$, compared to the evolution of $\langle c \rangle$ computed from the high-fidelity model at $t = 0.5,\ 1.5$, for an anomalous scenario.

For the scenarios with the largest $\langle u'^2 \rangle^{1/2} / \langle u \rangle$ and ℓ / L_x, exhibiting the most anomalous diffusion, the push-forward exhibited more substantial differences from the true evolution of $\langle c \rangle$, failing to capture the location of the bulk (or peak) of the concentration profile, as shown in Figure 5.10. Compared to the previous case, the 95% confidence interval of the eigenvalues does not encapsulate the mean for the lowest wavenumbers, as shown in Figure 5.12. The mean imaginary parts of $\widetilde{\boldsymbol{\lambda}}$ for the stochastic linear representation are uniformly smaller than the true values. As discussed in Section 4.3 and shown in (4.8), the advection velocity for each Fourier coefficient $\langle \hat{c}_k \rangle_y$ is $\langle u \rangle - \Im[\widetilde{\lambda}_k] / a_k$, where a_k is the wavenumber. Since $\Im[\widetilde{\lambda}_k]$

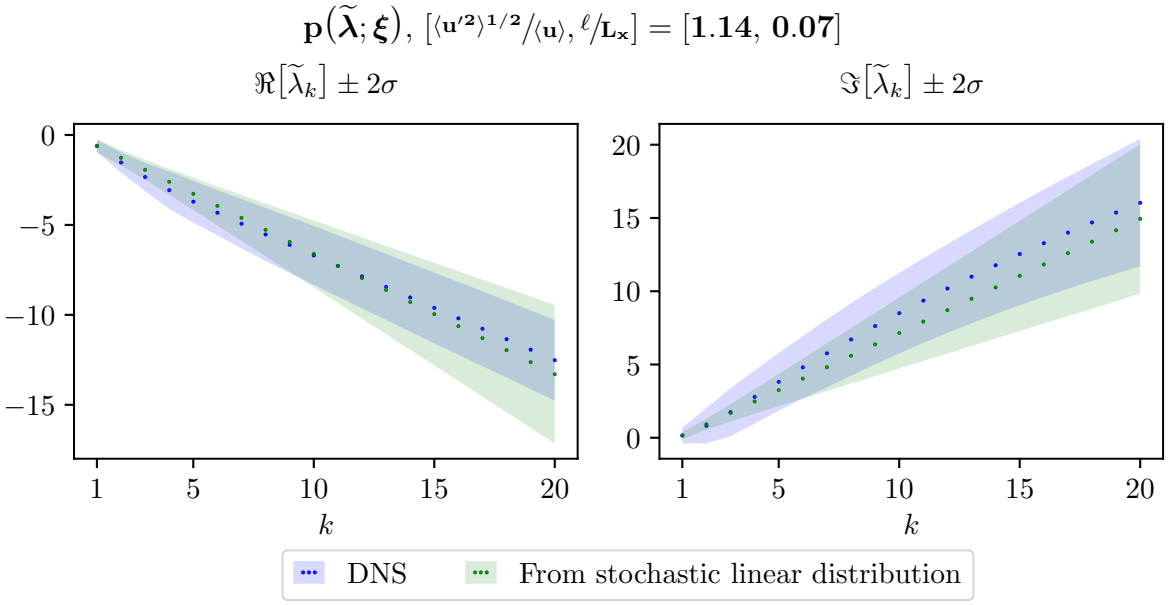

Figure 5.8: The summary statistics of the real and imaginary parts of $\widetilde{\boldsymbol{\lambda}}$, computed directly and from the approximate stochastic distribution $p(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi})$, for an anomalous scenario.

are positive, larger values correspond to slower advection velocities. By under-predicting the imaginary parts of $\widetilde{\boldsymbol{\lambda}}$, the stochastic linear representation over-predicts the advection velocity across a range of scales. This explains why the peak of the push-forward of $\langle c \rangle$ is farther downstream than the true evolution. However, even for this most extreme case considered in this study, the true mean's value at the downstream endpoint was encapsulated in the high-probability region of the push-forward, as shown in Figure 5.13. This suggests that even though the effect of dispersion is not captured perfectly by this simple model-form uncertainty representation, it may still be adequate, depending on the quantity of interest in the problem.

On the other hand, for the smallest $\langle u'^2 \rangle^{1/2}/\langle u \rangle$ and ℓ/L_x, a scenario that does not exhibit anomalous diffusion, the 95% confidence interval of the push-forward does not significantly overlap with the true mean, as shown in Figure 5.14. Furthermore, as shown in Figure 5.15, the push-forward predicts negative concentrations at the downstream endpoint L_x as a function of time. This indicates that although enforcing covariance mitigated the

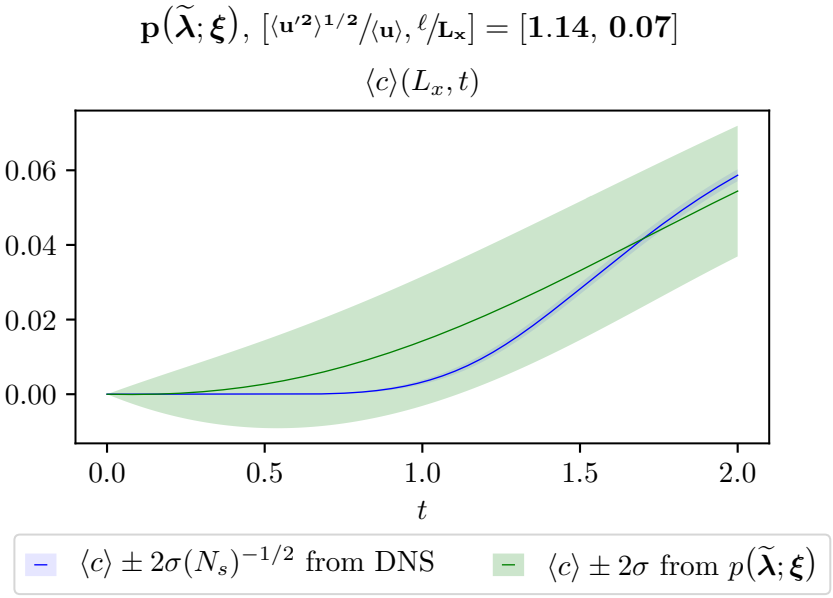

Figure 5.9: The push-forward of the stochastic formulation of $p(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi})$ to $\langle c \rangle$, compared to the evolution of $\langle c \rangle$ computed from the high-fidelity model at the downstream endpoint, L_x, for an anomalous scenario.

issue of oscillations and negative concentrations, further constraints must be derived to guarantee positivity is preserved. The push-forward exhibits more diffusion in $\langle c \rangle$ than exists in the true mean. This is likely due to the linear model for the mean of $\Re[\widetilde{\boldsymbol{\lambda}}]$, which predicts larger negative real parts than the true eigenvalues, as shown in Figure 5.16. Future refinements of the current model-form uncertainty representation could address this by allowing for a higher-degree model of the mean as a function of k.

These results indicate that the current formulation is not applicable for scenarios near the lower and upper bounds of the scenario space, $\langle u'^2 \rangle^{1/2}/\langle u \rangle \times \ell/L_x \in [0.5, 1.5] \times [0.02, 0.1]$. Along these boundaries, the assumption that the mean of $p(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi})$ as a function of k varies linearly is not valid. Furthermore, although $p(\widetilde{\boldsymbol{\lambda}})$ was assumed to be independent of $\langle v'^2 \rangle^{1/2}/\langle u \rangle$, this was based on the studies performed in Chapter 4 for the range $\langle v'^2 \rangle^{1/2}/\langle u \rangle \in [0.5, 1]$. This assumption may be invalid for scenarios with $\langle v'^2 \rangle^{1/2}/\langle u \rangle$ outside this range.

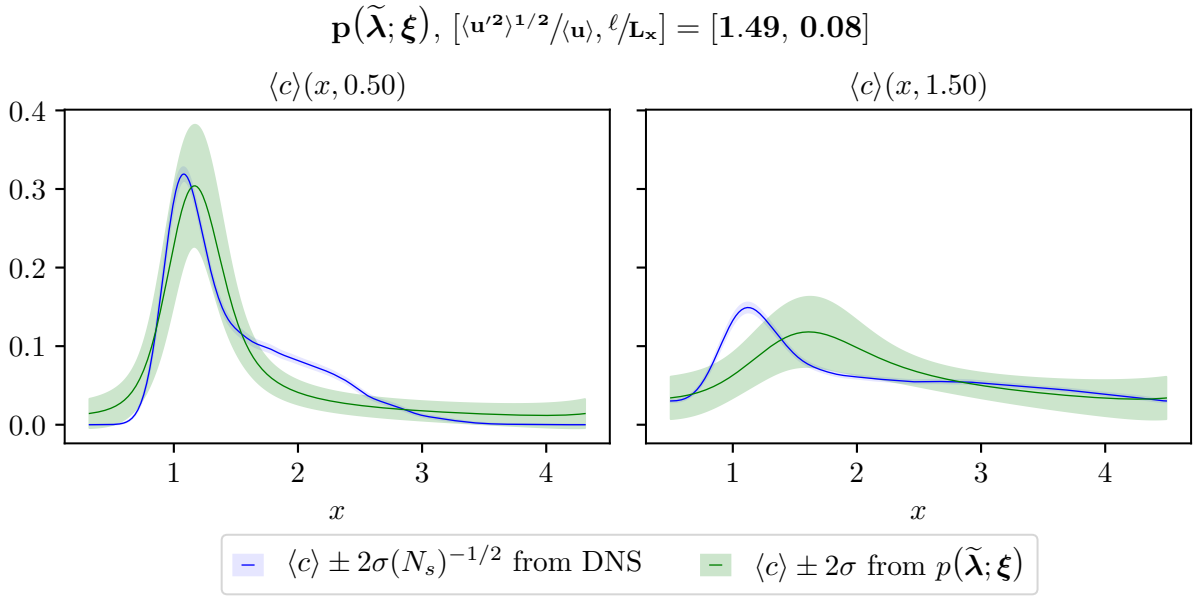

Figure 5.10: The push-forward of the stochastic formulation of $p(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi})$ to $\langle c \rangle$, compared to the evolution of $\langle c \rangle$ computed from the high-fidelity model at $t = 0.5,\ 1.5$, for an extremely anomalous scenario.

5.4 Conclusions

In summary, a stochastic formulation of model uncertainty was developed in terms of a stochastic linear operator $\widetilde{\mathcal{L}}$ acting on $\langle c \rangle_y$. The uncertainty in the model was represented in terms of $\widetilde{\mathcal{L}}$'s eigenvalues, $\widetilde{\boldsymbol{\lambda}}$. Deterministic physical constraints, scenario-dependence, and stochasticity were all imposed on a simple stochastic model of the eigenvalues' distribution, $p(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi})$, with uncertain hyperparameters $\boldsymbol{\xi}$. Quantities varying as a function of k were approximated with linear models, while its covariance was modeled with a rank-2 truncated eigendecomposition. The means and variances of the distributions of $\boldsymbol{\xi}$ were modeled with low-order polynomial fits to scenario parameters.

Although the current formulation is not valid in all scenarios, it shows promise as a basis for further improvements. Allowing for higher-order models of the mean behavior of $\widetilde{\lambda}_k$ as a function of k would improve predictions of $\langle c \rangle$. This would especially improve the

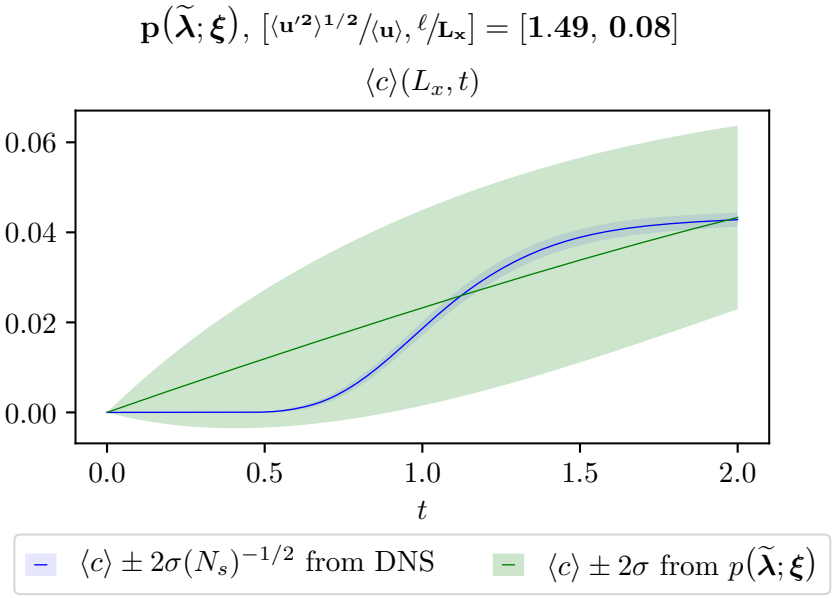

Figure 5.11: The push-forward of the stochastic formulation of $p(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi})$ to $\langle c \rangle$, compared to the evolution of $\langle c \rangle$ computed from the high-fidelity model at the downstream endpoint, L_x, for an extremely anomalous scenario.

representation of $p(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi})$ for extreme scenarios in terms of $\langle u'^2 \rangle^{1/2}/\langle u \rangle$ and ℓ/L_x, for which the mean behavior is nonlinear as a function of k. Although limiting arguments were made to constrain the functional fits of the mean and standard deviations of the hyperparameters on the lower end of the scenario space, no such arguments were made in the limit of large $\langle u'^2 \rangle^{1/2}/\langle u \rangle$ and ℓ/L_x. This would help the predictiveness of the uncertainty representation outside of the range of scenarios used to compute the polynomial fits. Across all scenarios, forward propagation of the directly-computed stationary values of $\widetilde{\boldsymbol{\lambda}}$ to a 95% confidence interval for $\langle c \rangle$ encompassed the true evolution of $\langle c \rangle$ from the high-fidelity model. This indicates that it is possible to derive a sufficiently rich formulation for $p(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi})$ to encompass the true evolution of $\langle c \rangle$.

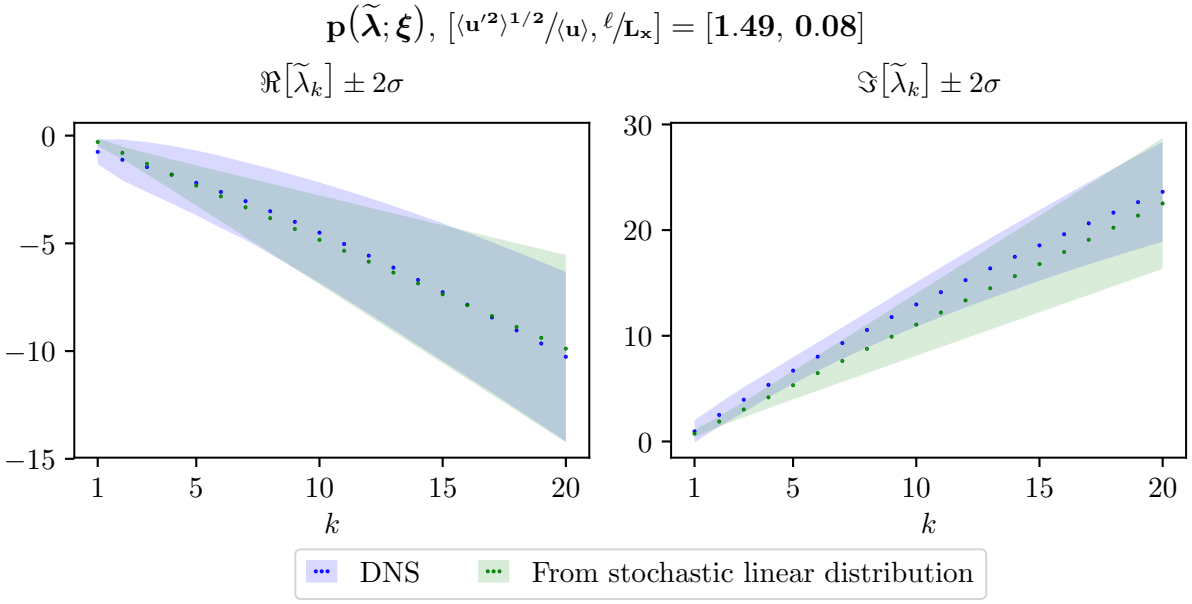

Figure 5.12: The summary statistics of the real and imaginary parts of $\widetilde{\boldsymbol{\lambda}}$, computed directly and from the approximate stochastic distribution $p(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi})$, for an extremely anomalous scenario.

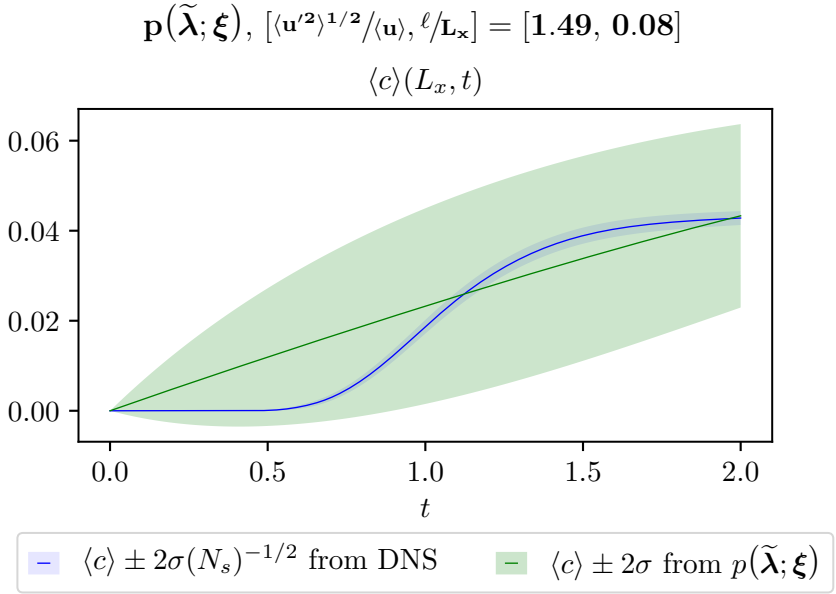

Figure 5.13: The push-forward of the stochastic formulation of $p(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi})$ to $\langle c \rangle$, compared to the evolution of $\langle c \rangle$ computed from the high-fidelity model at the downstream endpoint, L_x, for an extremely anomalous scenario.

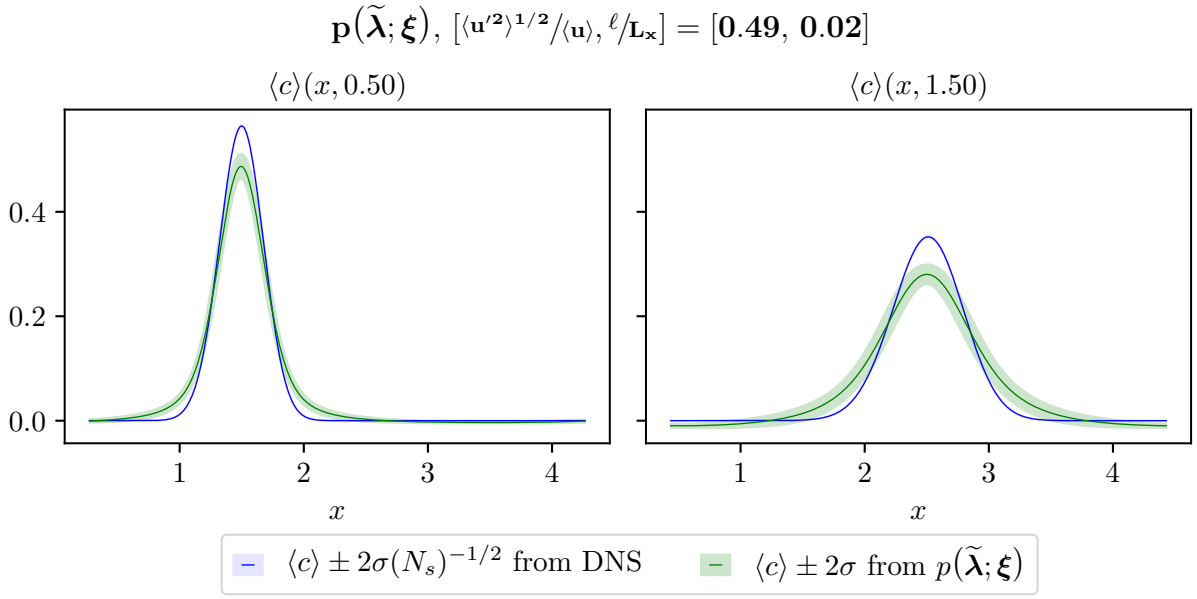

Figure 5.14: The push-forward of the stochastic formulation of $p(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi})$ to $\langle c \rangle$, compared to the evolution of $\langle c \rangle$ computed from the high-fidelity model at $t = 0.5,\ 1.5$, for a nonanomalous scenario.

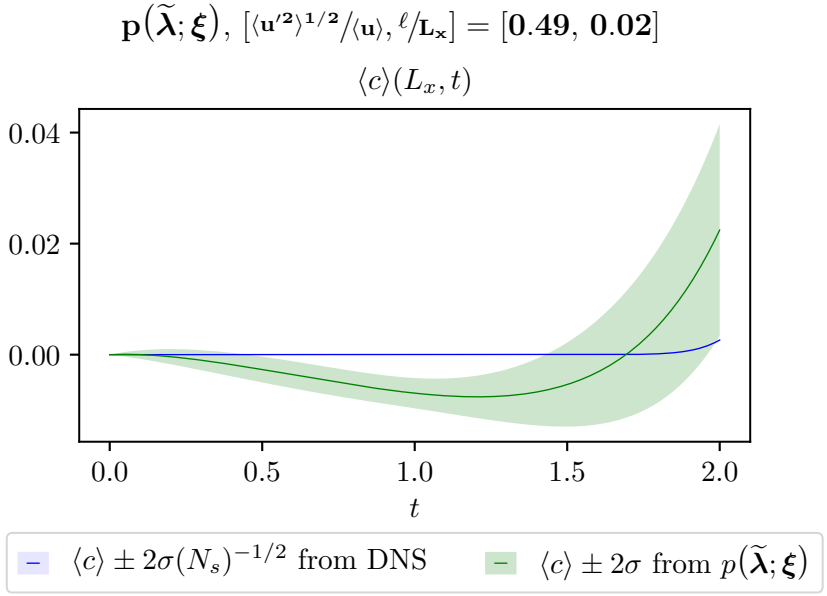

Figure 5.15: The push-forward of the stochastic formulation of $p(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi})$ to $\langle c \rangle$, compared to the evolution of $\langle c \rangle$ computed from the high-fidelity model at the downstream endpoint, L_x, for a nonanomalous scenario.

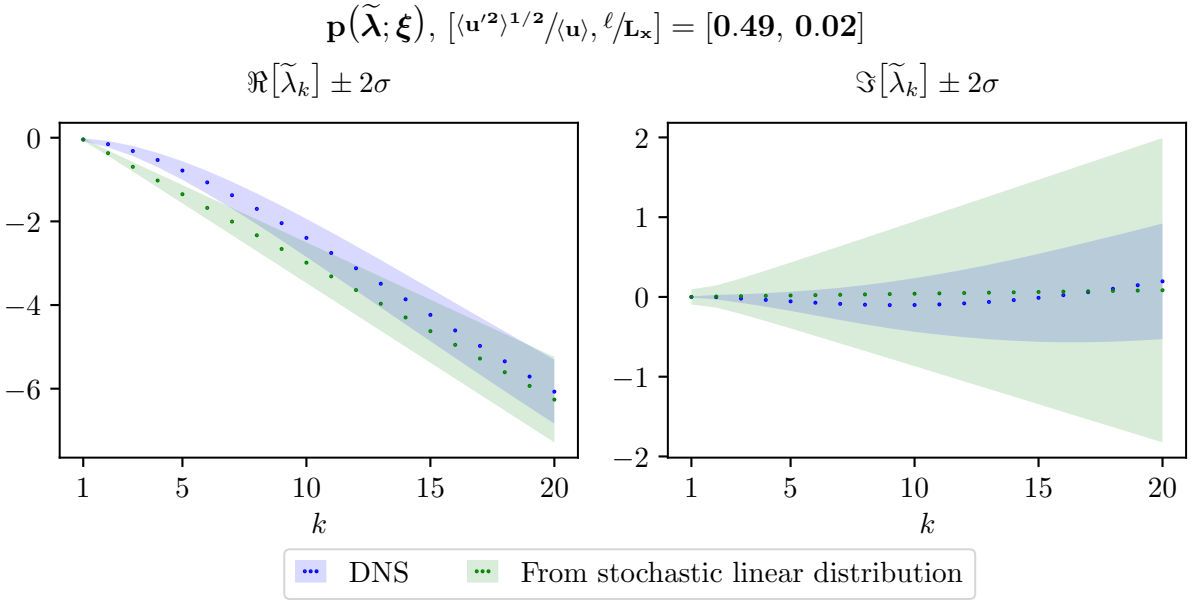

Figure 5.16: The summary statistics of the real and imaginary parts of $\widetilde{\boldsymbol{\lambda}}$, computed directly and from the approximate stochastic distribution $p(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi})$, for a nonanomalous scenario.

Chapter 6

Conclusions

The goal of this work was to explore two new avenues of representing model-form uncertainty. First, the challenges and feasibility of representing model-form uncertainty in multiscale problems arising from missing dependencies on small-scale dynamics were assessed. Second, a novel representation of model-form uncertainty using an infinite-dimensional stochastic operator was considered. These issues were explored in the context of a testbed problem, representing uncertainty in a field-scale model of mean contaminant transport through heterogeneous porous media.

To provide confidence in the representation for prediction, constraints based on physical arguments and scenario dependence were imposed on the operator. In Chapter 2 it was found that several physical constraints yielded simple mathematical requirements for the structure of the operator. For instance, linearity and shift-invariance of the governing equations required the same behavior in the operator. This resulted in its representation via its eigendecomposition and constrained its eigenfunctions to be the Fourier modes. However, a simple, constructive method of constraining the operator to preserve the positivity of the state could not be determined in the course of this work. This is likely an artifact of the eigenfunctions of the operator being the Fourier modes, which are not strictly positive, rather than an inherent issue with the infinite-dimensional operator formulation. Nevertheless, the challenge with enforcing positivity indicates that while many constraints are intuitive and

simple to enforce, some conditions will require more nuanced and complex approaches.

In Chapter 3, the feasibility of inferring the mean structure of the stochastic operator using observations of the mean state was explored. A Bayesian inverse problem was posed to infer the operator's eigenvalues. Even with precise and abundant data, the diffusive nature of the physical process limited the amount of information about the eigenvalues that could be recovered. Sparser, noisy data was not informative enough to penalize eigenvalues that produced nonphysical evolutions in the mean concentration, such as oscillations in the tails and negative values. As with Bayesian inference of infinite-dimensional fields, this study showed that a significant amount of prior information will be needed to ensure successful inference of an infinite-dimensional operator with finite data.

In Chapter 4, a novel method to determine the eigenvalues of the operator directly was introduced. The method exploits the high-fidelity model by setting its initial condition to be only one of the eigenfunctions of the stochastic operator appearing in the low-fidelity model. It also introduces a forcing that maintains the amplitude of the initially-excited eigenfunction for the duration of the evolution. The time-history of the evolved eigenfunction and the forcing are used to compute the corresponding time-history of the eigenvalue. This method allows for direct observation of sample eigenvalues and computation of summary statistics from the samples. This in turn enables study of their distribution, which encapsulates the uncertain dependence of the mean on the evolution of the microstate. These observations provided new insights into the macroscopic effects of the microstructural evolution. Furthermore, they were used to define the stochastic formulation of the uncertainty representation in terms of the probability distribution of the operator's eigenvalues.

In Chapter 5, a stochastic representation of the eigenvalues' distribution was constructed. Sensitivity analysis using different representations of the distribution helped identify the most important features to include in the final formulation. A relatively low-dimensional (40-parameter) stochastic representation employing linear models for wavenumber depen-

dence of the eigenvalues' mean and covariance and low-order polynomial fits for scenario dependence was used to predict the mean evolution of the concentration with uncertainty bounds. For scenarios exhibiting anomalous diffusion, the representation was remarkably successful in reproducing the mean evolution of $\langle c \rangle$ from the high-fidelity model. However, for scenarios that did not exhibit anomalous diffusion, it was not as successful. This is likely due to modeling the mean eigenvalues' dependence on wavenumber as linear, when closer to quadratic behavior was observed for such scenarios in Chapter 4.

To improve the predictive ability of the uncertainty representation, refinements to the current formulation could be made. For example, more complex models of the wavenumber dependence of the operator's eigenvalues would likely improve its predictions in the limiting case where microscale dynamics do not affect the macroscale. Limiting arguments in terms of scenario could further constrain the polynomial fits in the distribution. For the purposes of this work, however, the predictive capability of even this simple formulation is promising. Furthermore, forward propagation of stationary values of directly-computed eigenvalue ensembles encompassed the true evolution of the mean across all scenarios in the study. This indicates that a rich enough approximation of the eigenvalues' distribution would successfully encompass the true evolution of the mean concentration.

The testbed problem was an idealized version of the true problem of representing contaminant transport through heterogeneous porous media. There are several potential avenues to extend the current work. First, the testbed problem focused on representing the evolution of the mean concentration only. Future work focused on representing the evolution of the variance of the concentration would enable more realistic predictions regarding the probable evolution in any particular scenario. Second, a two-dimensional representation of the dynamics was assumed. Varying dependence on wavenumber between real and imaginary parts of the mean eigenvalues and with increasing wavenumber were observed in Chapter 4 for the two-dimensional problem. Applying the methodology described in Chapter 4 to a three-

dimensional version of the problem would test the veracity of the observations. Currently popular models for dispersion, gradient-diffusion and fractional-derivative models, assume a fixed dependence on wavenumber, as well as the same dependence between real and imaginary parts of the mean eigenvalues. If the observed complex dependence on wavenumber persists in three dimensions, it would indicate that these popular models are inconsistent with the observed dynamics of the dispersion. Finally, the assumption of homogeneous statistics for the permeability field can be relaxed to allow for mean statistics that vary slowly across the computational domain. The current formulation can be extended using a WKB approximation of the statistics, maintaining an assumption of homogeneity for rapidly-varying media characteristics and additionally representing the slowly-varying part as in [37].

This study explored the feasibility of representing model-form uncertainty caused by an uncertain dependence on small-scale dynamics. The testbed problem exhibited a significant amount of structure. Linearity in $\langle c \rangle$, statistical homogeneity, and the diffusive nature of the problem were exploited to derive a tractable, low-dimensional representation of the uncertain infinite-dimensional dependence. Linearity is not a common feature of multiscale problems. However, for nonlinear problems, the stochastic linear operator formulation described here can serve as a first-order representation of model-form uncertainty in a linearization about the mean. Many multiscale models assume statistical homogeneity at small scales, e.g. multiscale material models. For models with this property, the formulation of the stochastic operator and the methodology of directly observing its eigenvalues can serve as a basis for extending to other applications.

This work shows that an infinite-dimensional stochastic operator is a powerful approach for representing model-form uncertainty. The representation of the operator with respect to its eigendecomposition yielded a tractable means of constraining its behavior to respect physical constraints. Furthermore, understanding an operator's action through the lens of its eigendecomposition is both intuitive and well grounded in mathematical theory. Though

a simple, constructive constraint to ensure the operator preserves positivity could not be found, this challenge is an artifact of the eigenfunctions of the problem being the Fourier modes, rather than an inherent limitation of the infinite-dimensional operator representation. Overall, the formulation enabled a wide range of prior information to be encoded in the operator's structure. It also enabled a novel, intrusive study of the uncertain dependence in the problem using the methodology described in Chapter 4. The success and intuitiveness of a stochastic operator formulation both here and in [13], for two different application problems, illustrates the promise of this approach to representing model-form uncertainty.

Appendix A

Detailed advection-diffusion equation implementation

This appendix describes the discretization scheme used for computing the evolution of a single Fourier mode through the detailed 2D advection-diffusion equation. The system of equations to be evolved is

$$\begin{aligned} \frac{\partial c}{\partial t} + \mathbf{u} \cdot \nabla c = \nu_p \Delta c + f, \quad & x \in (0, L_x),\ y \in (0, L_y) \\ c(0, y) = c(L_x, y), \quad & y \in [0, L_y] \\ \frac{\partial c}{\partial y} = 0, \quad & x \in [0, L_x], y = 0, L_y, \\ c(x, y, t) = c_0(x, y). & \end{aligned} \tag{A.1}$$

The forcing function f is nonzero only in the case when the initial condition is a single Fourier mode as is required for the computational spectroscopy method. The velocity $\mathbf{u}$ is incompressible, and the domain lengths are $L_x = 4, L_y = 1$ unless otherwise specified.

A.1 Spatial discretization

The solution is periodic in the streamwise direction, so a Fourier discretization is employed in x:

$$c(x, y) \approx \sum_k \hat{c}_k(y) \exp(i a_k x).$$

The data for the physical domain and the Fourier coefficients will be arranged as

$$\mathbf{C} = \begin{bmatrix} c(x_0, y_0) & c(x_1, y_0) & \cdots \\ c(x_0, y_1) & c(x_1, y_1) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}, \quad \hat{\mathbf{C}} = \begin{bmatrix} \hat{c}_0(y_0) & \hat{c}_1(y_0) & \cdots \\ \hat{c}_0(y_1) & \hat{c}_1(y_1) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}.$$

The solution to the system may be complex, e.g. when the initial condition is a Fourier mode, so both $\mathbf{C}$ and $\hat{\mathbf{C}}$ are complex matrices in $\mathbb{C}^{N_y \times N_x}$. The grid is square with $N_x, N_y = 128$ per unit length, determined by a mesh-convergence study. Then for the domain lengths $L_x = 4, L_y = 1,\ N_x = 512,\ N_y = 128$.

A B-Spline collocation scheme is employed in the depthwise direction [38]. The Fourier coefficients are written as a linear combination of nth order B-splines (7^{th} order splines were used in this work):

$$\hat{c}_k(y_m) = \sum_j^{N_y} c_j^k B_j^n(y_m).$$

B_j^n can be determined by the de Boor recursive relations detailed in [38], and y_m are the Greville abscissae [39], which are defined in terms of uniform knot points. Here c_j^k is the jth coefficient in the B-spline expansion of the Fourier coefficient $\hat{c}_k$. The mth order derivative of a B-spline is defined by another recursive relationship. These are used to define matrix expressions for the expansion and derivatives of the expansion at collocation points y_m:

$$\begin{aligned} \hat{\mathbf{c}}_\mathbf{k} &= B_0 \mathbf{c}^\mathbf{k}, & \mathbf{c}^\mathbf{k} &= B_0^{-1} \hat{\mathbf{c}}_\mathbf{k} \\ \frac{\partial \hat{\mathbf{c}}_\mathbf{k}}{\partial y} &= B_1 \mathbf{c}^\mathbf{k}, & \frac{\partial^2 \hat{\mathbf{c}}_\mathbf{k}}{\partial y^2} &= B_2 \mathbf{c}^\mathbf{k}, \end{aligned} \tag{A.2}$$

where matrices B_0, B_1, B_2 are constructed according to the recursion relations using the C library GSL [40]. These matrices are computed at the beginning of the simulation.

A fully spectral solution in the x direction is not an option in this case because the velocity

varies across the domain. A pseudospectral scheme is employed where ∇c is computed in wavespace and transformed to realspace to multiply with $\mathbf{u}$. A 3/2 padding is performed for c and $\mathbf{u}$ to avoid aliasing [41]. Given this discretization, the gradient is computed and passed to realspace with the following relation:

$$\begin{aligned}\frac{\partial \mathbf{C}}{\partial x} &= \mathrm{DFT}^{-1}\left(\begin{bmatrix} \text{—} & i\mathbf{a_k} & \text{—} \\ \text{—} & i\mathbf{a_k} & \text{—} \\ & \vdots & \end{bmatrix} \odot \begin{bmatrix} \text{—} & \hat{\mathbf{C}}[0,:] & \text{—} \\ \text{—} & \hat{\mathbf{C}}[1,:] & \text{—} \\ & \vdots & \end{bmatrix}\right), \\ \frac{\partial \mathbf{C}}{\partial y} &= \mathrm{DFT}^{-1}[B_1 \mathbf{C_B}],\end{aligned} \tag{A.3}$$

where $\odot$ denotes element-wise multiplication.

The Laplacian is computed by

$$\Delta\hat{\mathbf{C}} = \begin{bmatrix} \text{—} & (-\mathbf{k_x^2}) & \text{—} \\ \text{—} & (-\mathbf{k_x^2}) & \text{—} \\ & \vdots & \end{bmatrix} \odot \begin{bmatrix} \text{—} & \hat{\mathbf{C}}[0,:] & \text{—} \\ \text{—} & \hat{\mathbf{C}}[1,:] & \text{—} \\ & \vdots & \end{bmatrix} + B_2 \mathbf{C_B}.$$

FFTW is used for the FFTs [42]. The inversion of the B-spline matrices to obtain B-spline coefficients is performed using the linear algebra methods developed by Myoungkyu Lee, detailed in [43]. B-spline matrices are banded but not sparse, so storing the full matrix is inefficient. Lee's linear algebra methods store the matrices in a packed form that only keep nonzero entries. The methods also feature customized inverse solves that account for the packed structure and are optimized for performance. The spline coefficients are computed one wavenumber at a time, so computed the entire $\mathbf{C_B}$ requires looping over all k.

A.1.1 Computing depthwise averages

All samples for the computational spectroscopy process are depthwise averaged to minimize storage. This is done using B-splines as follows. Recall that the n^{th}-order B-spline interpolation is written

$$f(y) \approx \sum_{j}^{N_y} f_j^B B_j^n(y).$$

Then the y average is

$$\begin{aligned}\frac{1}{L_y}\int_0^{L_y} f(y) &\approx \frac{1}{L_y}\sum_{j}^{N_y} f_j^B \int_0^{L_y} B_j^n(y) \\ &= \frac{1}{L_y}\sum_{j}^{N_y} f_j^B w_j \\ &= \frac{1}{L_y}(\mathbf{w}, \mathbf{f^B})\end{aligned}$$

where the integral of the B-splines is denoted w_j for "weights" and $(\cdot,\cdot)$ indicates an inner product.

It is also possible to compute the y average of a field in realspace without computing its spline coefficients. Let $\mathbf{f}$ denote the vector of evaluations of f at the collocation points, and let $\mathbf{f^B}$ denote its corresponding spline coefficients. The goal is to find a vector $\mathbf{v}$ such that $(\mathbf{v}, \mathbf{f}) \equiv (\mathbf{w}, \mathbf{f^B})$ The collocation relations require that $\mathbf{f} = B_0\mathbf{f^B}$, so

$$(\mathbf{w}, \mathbf{f^B}) = (\mathbf{w}, B_0^{-1}\mathbf{f}) = (B_0^{-T}\mathbf{w}, \mathbf{f}) \implies \mathbf{v} \equiv B_0^{-T}\mathbf{w}.$$

The integration weights $\mathbf{v}, \mathbf{w}$ are computed during setup along with the spline matrices B_0, B_1, B_2.

A.2 ADE Implementation

Given a velocity field and an initial condition by the ensemble average code, the ADE section of the code will need to generate a time-history of the depthwise-averaged Fourier coefficients and other y moments. To do this, it will march forward in time with the 2D solution and take the depthwise averages at a specified interval. These $y-$averaged solutions are stored in matrices are saved to disk for postprocessing.

A.2.1 Forcing derivation for computational spectroscopy

The detailed system of equations for the computational spectroscopy is

$$\begin{aligned}
\frac{\partial c}{\partial t} + \mathbf{u}\cdot\nabla c = \nu_p \Delta c + f, \quad x \in (0, L_x), \quad y \in (0, L_y) \\
c(0,y) = c(L_x, y), \quad y \in (0, L_y) \\
\frac{\partial c}{\partial y} = 0, \quad y = 0, L_y, \quad x \in [0, L_x] \\
c_0(x,y) = \exp\left(i a_k x\right),
\end{aligned}$$

where $f = \alpha(t)\left\langle \hat{c}_k \right\rangle_y,\ \alpha \in \mathbb{R}$. The y-averaged evolution equation is

$$\frac{\partial \left\langle c \right\rangle_y}{\partial t} + \frac{\partial \left\langle \mathbf{u}c \right\rangle_y}{\partial x} = \nu_p \frac{\partial^2 \left\langle c \right\rangle_y}{\partial x^2} + \alpha \left\langle \hat{c}_k \right\rangle_y .$$

In wavespace this yields

$$\begin{aligned}
\frac{\mathrm{d}\left\langle \hat{c}_k \right\rangle_y}{\mathrm{d}t} = \hat{R}_k + \alpha \left\langle \hat{c}_k \right\rangle_y , \\
\hat{R}_k \equiv -\nu_p (a_k)^2 \left\langle \hat{c}_k \right\rangle_y - (i a_k) \left\langle \widehat{(\mathbf{u}c)}_k \right\rangle_y .
\end{aligned}$$

The goal is to find an $\alpha(t)$ such that $\langle \hat{c}_k \rangle_y = 1 \; \forall t$, or, equivalently, $\partial_t \left| \langle \hat{c}_k \rangle_y \right|^2 = 0$. Then

$$\begin{aligned}
\frac{\mathrm{d}\left| \langle \hat{c}_k \rangle_y \right|^2}{\mathrm{d}t} &= \frac{\mathrm{d}\left\langle \hat{c}_k \right\rangle_y^* \langle \hat{c}_k \rangle_y}{\mathrm{d}t} = \langle \hat{c}_k \rangle_y^* \frac{\mathrm{d}\left\langle \hat{c}_k \right\rangle_y}{\mathrm{d}t} + \frac{\mathrm{d}\left\langle \hat{c}_k \right\rangle_y^*}{\mathrm{d}t} \langle \hat{c}_k \rangle_y \\
&= \langle \hat{c}_k \rangle_y^* \left(\hat{R}_k + \alpha \langle \hat{c}_k \rangle_y \right) + \left(\hat{R}_k + \alpha \langle \hat{c}_k \rangle_y \right)^* \langle \hat{c}_k \rangle_y \\
&= 2\Re \left[\langle \hat{c}_k \rangle_y^* (\hat{R}_k + \alpha \langle \hat{c}_k \rangle_y) \right] \\
&= 2\Re \left[\langle \hat{c}_k \rangle_y^* \hat{R}_k \right] + 2\alpha \Re \left[\langle \hat{c}_k \rangle_y^* \langle \hat{c}_k \rangle_y \right] \\
&= 2\Re \left[\langle \hat{c}_k \rangle_y^* \hat{R}_k \right] + 2\alpha \left| \langle \hat{c}_k \rangle_y \right|^2 .
\end{aligned}$$

To satisfy the constant time derivative of the norm of the coefficient, α must satisfy

$$\alpha = \frac{-\Re \left[\langle \hat{c}_k \rangle_y^* \hat{R}_k \right]}{\left| \langle \hat{c}_k \rangle_y \right|^2} = \frac{-\Re \left[\langle \hat{c}_k \rangle_y^* \left(-\nu_p (a_k)^2 \langle \hat{c}_k \rangle_y - (ia_k) \left\langle \widehat{(\mathbf{u}c)}_k \right\rangle_y \right) \right]}{\left| \langle \hat{c}_k \rangle_y \right|^2}.$$

A.2.2 Timestepping scheme

Taking the most general form of the ADE that includes the forcing function, the analytical equation is

$$\frac{\partial c}{\partial t} = \nu_p \Delta c - \mathbf{u} \cdot \nabla c + f$$

The ADE is advanced using the low-storage, implicit-explicit RK3 scheme developed by Spalart, Moser and Rogers in [44], herein called the SMR scheme. Following the notation

laid out in their paper, the ADE is split into a linear and nonlinear part as

$$\begin{aligned}\frac{\partial c}{\partial t} &= \mathcal{L}(c) + \mathcal{N}(c),\\ \mathcal{L}(c) &= \nu_p \Delta c,\\ \mathcal{N}(c) &= -\mathbf{u}\cdot\nabla c + f.\end{aligned}$$

The linear terms are treated implicitly, while the nonlinear terms are treated explicitly. Let $c_n = c(x, y, t_n)$. The three-step scheme takes the form

$$\begin{aligned}c' &= c_n + \Delta t\left[\mathcal{L}(\alpha_1 c_n + \beta_1 c') + \gamma_1 \mathcal{N}(c_n)\right]\\ c'' &= c' + \Delta t\left[\mathcal{L}(\alpha_2 c' + \beta_2 c'') + \gamma_2 \mathcal{N}(c') + \zeta_1 \mathcal{N}(c_n)\right]\\ c_{n+1} &= c'' + \Delta t\left[\mathcal{L}(\alpha_3 c'' + \beta_3 c_{n+1}) + \gamma_3 \mathcal{N}(c'') + \zeta_2 \mathcal{N}(c'),\right],\end{aligned}$$

$$\begin{aligned}\alpha_1 &= \frac{29}{96}, & \beta_1 &= \frac{37}{160}, & \gamma_1 &= \frac{8}{15}, & \zeta_1 &= -\frac{17}{60},\\ \alpha_2 &= -\frac{3}{40}, & \beta_2 &= \frac{5}{24}, & \gamma_2 &= \frac{5}{12}, & \zeta_2 &= -\frac{5}{12}.\\ \alpha_3 &= \frac{1}{6} & \beta_3 &= \frac{1}{6} & \gamma_3 &= \frac{3}{4}\end{aligned}$$

Substituting the Fourier/B-spline discretization of c and denoting the discrete versions of $\mathcal{L}, \mathcal{N}$ as $\widehat{\mathcal{N}}, \widehat{\mathcal{L}}$ into the scheme yields

$$\begin{aligned}B_0\mathbf{C_B}' &= B_0\mathbf{C_B}_n + \Delta t\left[\widehat{\mathcal{L}}(\alpha_1 \mathbf{C_B}_n + \beta_1 \mathbf{C_B}') + \gamma_1 \widehat{\mathcal{N}}(\mathbf{C_B}_n)\right]\\ B_0\mathbf{C_B}'' &= B_0\mathbf{C_B}' + \Delta t\left[\widehat{\mathcal{L}}(\alpha_2 \mathbf{C_B}' + \beta_2 \mathbf{C_B}'') + \gamma_2 \widehat{\mathcal{N}}(\mathbf{C_B}') + \zeta_1 \widehat{\mathcal{N}}(\mathbf{C_B}_n)\right]\\ B_0\mathbf{C_B}_{n+1} &= B_0\mathbf{C_B}'' + \Delta t\left[\widehat{\mathcal{L}}(\alpha_3 \mathbf{C_B}'' + \beta_3 \mathbf{C_B}_{n+1}) + \gamma_3 \widehat{\mathcal{N}}(\mathbf{C_B}'') + \zeta_2 \widehat{\mathcal{N}}(\mathbf{C_B}')\right].\end{aligned}$$

Rearranging so that all terms related to the advanced spline coefficients are on the left-hand

side of the equations gives

$$\begin{aligned}
\left[B_0 - \Delta t\beta_1\widehat{\mathcal{L}}\right]\mathbf{C_B}' &= \left[B_0 + \Delta t\alpha_1\widehat{\mathcal{L}}\right]\mathbf{C}_{\mathbf{B}n} + \Delta t\gamma_1\widehat{\mathcal{N}}(\mathbf{C}_{\mathbf{B}n}) \\
\left[B_0 - \Delta t\beta_2\widehat{\mathcal{L}}\right]\mathbf{C_B}'' &= \left[B_0 + \Delta t\alpha_2\widehat{\mathcal{L}}\right]\mathbf{C_B}' + \Delta t\left[\gamma_2\widehat{\mathcal{N}}(\mathbf{C_B}') + \zeta_1\widehat{\mathcal{N}}(\mathbf{C}_{\mathbf{B}n})\right] \\
\left[B_0 - \Delta t\beta_3\widehat{\mathcal{L}}\right]\mathbf{C}_{\mathbf{B}n+1} &= \left[B_0 + \Delta t\alpha_3\widehat{\mathcal{L}}\right]\mathbf{C_B}'' + \Delta t\left[\gamma_3\widehat{\mathcal{N}}(\mathbf{C_B}'') + \zeta_2\widehat{\mathcal{N}}(\mathbf{C_B}')\right]
\end{aligned}$$

The right-hand side of each equation is computed with current and previous substep evaluations. Let the k^{th} column of $\mathbf{C_B}$ be denoted $\mathbf{c_B^k}$. Then, expanding out $\widehat{\mathcal{L}}$, the left-hand side for each substep can be written

$$\left[B_0 - \Delta t\beta_i\widehat{\mathcal{L}}\right]\hat{\mathbf{c}}_{\mathbf{B}}^{\mathbf{k}} = \left[B_0 - \Delta t\beta_i\left(-\nu_p a_k^2 B_0 + \nu_p B_2\right)\right]\mathbf{c_B^k}.$$

To take a substep the matrix is formed and inverted for each column, since a_k will vary. To ensure the zero-Neumann boundary condition is enforced at all substeps, the top and bottom of the mass matrix are set to the top and bottom rows of the discrete differential operator in y, B_1, and the right-hand side of the equation is set to zero.

A.2.3 Stability criteria for step size

Studies on stability criteria for the SMR scheme with a Fourier-Galerkin/B-Spline collocation spatial discretization are reported in [45–47]. The criteria are based on model convection and diffusion problems and result in the requirements

$$\Delta t_{\text{convective}} \leq \frac{|\lambda_I \Delta t|_{\max}}{|u|_\infty \lambda_x^{(1)} + |v|_\infty \lambda_y^{(1)}},$$

$$\Delta t_{\text{diffusive}} \leq \frac{|\lambda_R \Delta t|_{\max}}{\nu_p(\lambda_x^{(2)} + \lambda_y^{(2)})},$$

where $\lambda_x^{(1)}$ is the maximum imaginary eigenvalue magnitude for numerical ∂_x, $\lambda_y^{(1)}$ for numerical ∂_y, $\lambda_x^{(2)}$ for numerical ∂_{xx}, $\lambda_y^{(2)}$ for numerical ∂_{yy}. Finally, $|\lambda_I \Delta t|_{\max}$ and $|\lambda_R \Delta t|_{\max}$ are the maximum imaginary and real eigenvalues estimated from the SMR scheme's stability plot. Their values are

$$|\lambda_I \Delta t|_{\max} = \sqrt{3},$$
$$|\lambda_R \Delta t|_{\max} = 2.512.$$

In the homogeneous direction $\lambda_x^{(1)}, \lambda_x^{(2)}$ are clearly

$$\lambda_x^{(1)} = \frac{2\pi}{L_x}\frac{N_x}{2} = \frac{\pi N_x}{L_x} = \frac{\pi}{\Delta x}, \qquad \lambda_x^{(2)} = \left(\frac{\pi N_x}{L_x}\right)^2 = \left(\frac{\pi}{\Delta x}\right)^2.$$

The values of $\lambda_y^{(1)}, \lambda_y^{(2)}$ are less obvious, but numerical experiments showed that it is reasonable in this case to use the eigenvalues of the Fourier derivative as approximations of the true eigenvalues of the numerical derivative operator in the y direction. In that case, the values are

$$\lambda_y^{(1)} = \frac{\pi}{\Delta y}, \qquad \lambda_y^{(2)} = \left(\frac{\pi}{\Delta y}\right)^2,$$

where $\Delta y = L_y/N_y$, the distance between the uniform knots for the B-Spline discretization.

Because the viscous term in the ADE is advanced with an implicit method in the implicit-explicit scheme, it is unconditionally stable. For this reason, only the convective stability requirement is used to determine the step size for the scheme.

Appendix B

Pressure formulation

The pressure solve was formulated and implemented by Drs. Damon McDougall and Manav Vohra. It is detailed here for completeness.

B.1 Formulation

The pressure is governed by Darcy's law and an incompressibility constraint on $\mathbf{u}$.

$$\mathbf{u} = -\kappa \nabla p$$

$$\nabla \cdot \mathbf{u} = 0.$$

Composing these equations results in a variable-coefficient Poisson problem. Boundary conditions on p are derived from the boundary conditions on $\mathbf{u}$ that are required for the advection-diffusion system. Both components of the velocity must be periodic in x, and a no-flow condition at the top and bottom of the domain requires the vertical component v to satisfy zero-Dirichlet boundary conditions at $y = 0, L_y$. The corresponding boundary conditions will be on the partial derivatives of p through Darcy's Law. Then the system for

the pressure is

$$\nabla \cdot (\kappa \nabla p) = 0, \qquad x \in (0, L_x), y \in (0, L_y), \tag{B.1}$$

$$\frac{\partial p}{\partial y} = 0, \qquad x \in [0, L_x], y = 0, L_y, \tag{B.2}$$

$$\left(\frac{\partial p}{\partial x}\right)_{x=0} = \left(\frac{\partial p}{\partial x}\right)_{x=L_x}, \qquad y \in [0, L_y] \tag{B.3}$$

$$\left(\frac{\partial p}{\partial y}\right)_{x=0} = \left(\frac{\partial p}{\partial y}\right)_{x=L_x}, \qquad y \in [0, L_y]. \tag{B.4}$$

This system is ill-posed; adding any constant to a solution will yield another solution. Furthermore, enforcing periodic constraints on the gradient of p can be challenging. To solve this, p is written as

$$p(x, y) = p_0 + p_1 x + p_2 y + \tilde{p}(x, y).$$

The choice of p_0 does not affect the resulting flow field, so without loss of generality $p_0 = 0$. The goal is to reformulate the problem so that periodic boundary conditions are enforced on $\tilde{p}$ directly rather than on ∇p. The zero-Neumann boundary condition will be passed to $\tilde{p}$, so for both $\tilde{p}$ and p to satisfy the condition at the boundary it is required that $p_2 = 0$. The expression for p simplifies to

$$p(x, y) = p_1 x + \tilde{p}(x, y). \tag{B.5}$$

To show that $\tilde{p}$ is periodic, note that

$$p(0, y) - p(0, 0) = \int_0^y \frac{\partial p}{\partial y}(0, y')\mathrm{d}y'$$

$$p(L_x, y) - p(L_x, 0) = \int_0^y \frac{\partial p}{\partial y}(L_x, y')\mathrm{d}y'.$$

The integrands are equal because of (B.4), so

$$\begin{aligned} p(0,y) - p(0,0) &= p(L_x,y) - p(L_x,0) \\ \implies p(L_x,y) - p(0,y) &= p(0,0) - p(L_x,0) = C \quad \forall y. \end{aligned}$$

Substituting (B.5) into $p(L_x,y) - p(0,y) = C$ yields

$$p_1 L_x + \tilde{p}(L_x,y) - \tilde{p}(0,y) = C. \tag{B.6}$$

Clearly $p_1 = C/L_x$ if and only if $\tilde{p}$ is periodic. Substituting (B.5) into (B.1) yields

$$\begin{aligned} 0 &= \nabla \cdot (\kappa \nabla (p_1 x + \tilde{p}) \\ &= \nabla \cdot \begin{pmatrix} \kappa p_1 \\ \\ 0 \end{pmatrix} + \nabla \cdot (\kappa \nabla \tilde{p}) \\ &= p_1 \frac{\partial k}{\partial x} + \nabla \cdot (k \nabla \tilde{p}). \end{aligned}$$

Given p_1 the following system can be solved for $\tilde{p}$:

$$\begin{aligned} \nabla \cdot (k \nabla \tilde{p}) = -\frac{\partial k}{\partial x} p_1, &\quad x \in (0, L_x),\ y \in (0, L_y) \\ \frac{\partial \tilde{p}}{\partial y} = 0, &\quad x \in [0, L_x],\ y = 0, L_y \\ \tilde{p}(0,y) = \tilde{p}(L_x,y), &\quad y \in [0, L_y]. \end{aligned} \tag{B.7}$$

For a pressure drop (decreasing pressure with increasing x) p_1 should be negative. A unit pressure drop was used for this work, so $p_1 = -1$. Finally, the velocity can be derived from

$\tilde{p}$ and p_1 by

$$\begin{aligned} u &= -kp_1 - k\frac{\partial \tilde{p}}{\partial x} \\ v &= -k\frac{\partial \tilde{p}}{\partial y}. \end{aligned} \tag{B.8}$$

B.2 Implementation

The solution of system (B.7) was implemented in the finite element library FEniCS [48] using first-order Lagrange elements on a Cartesian grid. An iterative solve was required to invert the system, so the PETSc Krylov solver integrated into FEniCS was employed using the Conjugate Gradient method with Adaptive Multigrid (AMG) preconditioning. A finite element representation of the velocity was then derived from $\tilde{p}$ using (B.8) and used to evaluate the velocity on the mesh used for the advection-diffusion solve.

Appendix C

Divergence-free velocity projection

The velocity interpolated onto the grid for the advection-diffusion solver does not satisfy continuity with respect to the discrete differentiation operators used in the ADE, which is discretized using a Fourier-Galerkin/B-Spline-Collocation scheme as described in Appendix A. To find the nearest velocity in the L_2 sense that satisfies continuity with respect to the Fourier/B-spline differential operators, the finite element representation of the velocity evaluated on the ADE grid, denoted $\mathbf{u}$, is decomposed into a divergence-free and curl-free part:

$$\mathbf{u} = \mathbf{u_{df}} + \mathbf{u_{cf}}. \tag{C.1}$$

Then the divergence-free velocity can be computed by solving for $\mathbf{u_{cf}}$ with respect to the Fourier/B-spline operators and computing $\mathbf{u_{df}} = \mathbf{u} - \mathbf{u_{cf}}$. The two components of $\mathbf{u_{cf}} = [u_{cf}, v_{cf}]$ can be computed by solving the coupled system of equations obtained by taking the divergence and curl of (C.1) respectively:

$$\begin{aligned} \frac{\partial u_{cf}}{\partial x} + \frac{\partial v_{cf}}{\partial y} &= \nabla \cdot (\mathbf{u}), \\ -\frac{\partial u_{cf}}{\partial y} + \frac{\partial v_{cf}}{\partial x} &= 0. \end{aligned}$$

The decision to solve for the components of the curl-free part of the velocity rather than for a curl-free potential such that $\mathbf{u_{cf}} = \nabla\phi$ is based on the fact that the equations to

solve for the potential involve a second derivative ($\Delta\phi = \nabla \cdot (\mathbf{u})$). The discrete second derivative operator is not equivalent to applying two first derivative operators in a B-spline discretization, so the divergence-free projection using ϕ would not be as precise as solving a larger system for the individual components u_{cf}, v_{cf} that uses the first-derivative operator directly.

Analytically, the pressure satisfies a zero-Neumann boundary condition in the normal direction at the top and bottom of the domain, corresponding to a zero-Dirichlet boundary condition on v. However, the FEM implementation of the pressure solve only enforces this boundary condition weakly, so v is not exactly zero at $y = 0, L_y$. To ensure the boundary condition is respected for v_{df}, the Dirichlet condition $v_{cf} = v$ is enforced at the depthwise boundary.

Let $\hat{u}_k^{cf}$ and $\hat{v}_k^{cf}$ denote the k^{th} Fourier coefficient of u_{cf} and v_{cf} with respect to the Fourier-Galerkin discretization applied in the (homogeneous) x direction. Then the system of equations to be solved is

$$\begin{aligned} (ia_k)\hat{u}_k^{cf} + \frac{\partial \hat{v}_k^{cf}}{\partial y} &= \left(\widehat{\nabla \cdot (\mathbf{u})}\right)_k, \quad \forall k, y \in (0, L_y) \\ -\frac{\partial \hat{u}_k^{cf}}{\partial y} + (ia_k)\hat{v}_k^{cf} &= 0, \quad \forall k, y \in (0, L_y) \\ \hat{v}_k^{cf} &= \hat{v}_k, \quad \forall k, y = 0, L_y. \end{aligned} \tag{C.2}$$

Let the vector of B-spline coefficients representing the depthwise-dependent Fourier coefficients be denoted $\mathbf{u}_{\mathbf{B},\mathbf{k}}^{\mathbf{cf}}$ and $\mathbf{v}_{\mathbf{B},\mathbf{k}}^{\mathbf{cf}}$. Introducing the B-Spline discretization and following the notation introduced in Section A.1, the system of equations for the B-spline coefficients

is

$$
\begin{aligned}
(ia_k)B_0\mathbf{u}^{\mathbf{cf}}_{\mathbf{B},\mathbf{k}} + B_1\mathbf{v}^{\mathbf{cf}}_{\mathbf{B},\mathbf{k}} &= (\widehat{\nabla\cdot(\mathbf{u})})_k, \quad k \neq 0, y \in (0, L_y) \\
-B_1\mathbf{u}^{\mathbf{cf}}_{\mathbf{B},\mathbf{k}} + (ia_k)B_0\mathbf{v}^{\mathbf{cf}}_{\mathbf{B},\mathbf{k}} &= 0, \quad k \neq 0, y \in (0, L_y) \\
B_0\mathbf{v}^{\mathbf{cf}}_{\mathbf{B},\mathbf{k}} &= \hat{v}_k, \quad k \neq 0, y = 0, L_y.
\end{aligned}
$$

For $k = 0$, (C.2) simplifies to

$$
\begin{aligned}
\frac{\partial \hat{u}_0^{cf}}{\partial y} &= 0, \quad y \in (0, L_y), \\
\frac{\partial \hat{v}_0^{cf}}{\partial y} &= (\widehat{\nabla\cdot(\mathbf{u})})_0, \quad y \in (0, L_y), \\
\hat{v}_0^{cf} &= \hat{v}_0, \quad y = 0, L_y.
\end{aligned}
\tag{C.3}
$$

Then $\hat{u}_0^{cf}$ is constant in y and is set to 0 so that $\langle u \rangle = \langle u_{df} \rangle$. A smaller discrete system is solved to obtain the B-spline coefficients for $\hat{v}_0^{cf}$:

$$
\begin{aligned}
B_1\mathbf{v}^{\mathbf{cf}}_{\mathbf{B},\mathbf{0}} &= (\widehat{\nabla\cdot(\mathbf{u})})_0, y \in (0, L_y) \\
B_0\mathbf{v}^{\mathbf{cf}}_{\mathbf{B},\mathbf{0}} &= \hat{v}_0, \quad y = 0, L_y.
\end{aligned}
$$

The mass matrices can be inverted directly to solve for the spline coefficients of the curl-free component.

Appendix D

Statistics with averaging operator $\mathbb{E}\left[\langle\cdot\rangle_y\right]$

This appendix details the computation of the statistical quantities used for analysis in Chapter 4. Recall that the averaging operator in this work is a composition of an expectation and a depthwise average and is denoted $\langle\cdot\rangle = \mathbb{E}\left[\langle\cdot\rangle_y\right]_y$, while deviations from this mean are denoted $f' = f - \langle f\rangle$. Variances with respect to this operator are defined as would be expected, but care must be taken to compute the statistical variation of quantities computed with respect to this mean, which are needed for convergence analyses. The statistical quantities being computed for this work are detailed in Section D.1 and variance formulae are derived in Section D.2.

D.1 Statistics of interest

The quantities needed for a scenario-dependence study using computational spectroscopy are $\langle c\rangle$, $\langle c'^2\rangle$; $\langle u'^2\rangle_y$; $\langle v'^2\rangle_y$; $\langle u'c'\rangle_y$; $\left\langle \mathrm{Arg}\langle\hat{c}_k\rangle_y \right\rangle$, the phase of the $y-$averaged, forced coefficient; $\langle f\rangle$ and $\langle\alpha\rangle$, from the forcing term; and ℓ, the integrated autocorrelation length of u in the streamwise direction. All second-order centered moments are computed *a posteriori* using, e.g. $\langle c'^2\rangle = \langle c^2\rangle - \langle c\rangle^2$. The velocity statistics are additionally averaged in the streamwise (statistically homogeneous) direction to minimize sampling error.

The integrated autocorrelation length ℓ is a length scale and provides a measure of correlation in the streamwise direction. Because the velocity field is periodic, maximum decor-

relation is achieved at a half domain length, so the integration bounds of ℓ are defined to be from 0 to $L_x/2$:

$$\ell = \int_0^{L_x/2} \rho(x')\mathrm{d}x',$$
$$\rho(x') \equiv \frac{\langle u'(x)u'(x+x')\rangle}{\langle u'^2\rangle}.$$

The integrand ρ is computed as a function of x', then is numerically integrated. First, the numerator of ρ can be split into two parts using the Reynolds decomposition:

$$\rho(x')\left\langle u'^2\right\rangle = \langle u(x)u(x+x')\rangle - \langle u\rangle^2.$$

The mean $\langle u\rangle$ is computed independently, so only $\langle u(x)u(x+x')\rangle$ must be computed here. The statistical mean is computed by taking a sample average over all velocities in the ensemble as well as by averaging in the homogeneous (x) direction:

$$\langle u(x)u(x+x')\rangle \approx \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L_xL_y}\int_0^{L_y}\int_0^{L_x} u^{(i)}(x)u^{(i)}(x+x')\mathrm{d}x\mathrm{d}y,$$

where $u^{(i)}$ denotes the i^{th} velocity in the ensemble. The DFT is equivalent to discrete convolution, so the streamwise average is computed using the Fourier Convolution Theorem as follows:

$$\begin{aligned}\frac{1}{L_x}\int_0^{L_x} u(x)u(x+x')\mathrm{d}x &\approx \frac{1}{L_x}\sum_{i=0}^{N_x} u(x_i)u(x_{i+j})\Delta x, \quad \Delta x = \frac{L_x}{N_x}, \quad x' = j\Delta x \\ &= \frac{\Delta x}{L_x}\sum_{i=0}^{N_x} u(x_i)u(x_{i+j}) \\ &= \frac{1}{N_x}\mathrm{DFT}^{-1}[\hat{\mathbf{u}}\odot\hat{\mathbf{u}}^*],\end{aligned}$$

where $\odot$ represents a termwise multiplication and $(\cdot)^*$ represents conjugation.

D.2 Sample statistics and sampling error

The statistical variation of second-order centered statistics computed with respect to the composed averaging operator $\mathbb{E}\left[\langle\cdot\rangle_y\right]$ cannot be estimated using the variance from that mean. In this section, the sample variance for these second-order quantities is derived. Recall that a fluctuation from the mean $\mathbb{E}\left[\langle\cdot\rangle_y\right] \equiv \langle\cdot\rangle$ is denoted $f' \equiv f - \langle f\rangle$. A fluctuation from the statistical mean only will be denoted $f'^{\mathbb{E}} = f - \mathbb{E}[f]$

The standard statistical error analysis holds for all raw moments. For instance, the standard analysis holds for $\langle u\rangle_y, \langle c\rangle_y$ and $\left\langle \langle c\rangle_y'^2 \right\rangle$. It will not hold, however for $\langle u'c'\rangle$ or $\langle c'^2\rangle$, for example. The classical statistical analysis used to compute the mean and variance of a sampling distribution for a sample statistic will be outlined for a y-averaged quantity, then it will be extended to a second-order centered statistic with respect to $\langle\cdot\rangle$.

D.2.1 Sampling distribution of $y-$averaged quantities

The classical statistical analysis described here follows discussion in [49]. For simplicity of notation let $\sum_{i=1}^{N} \equiv \sum_i$. The goal is to determine the probability distribution of the sample statistic $\langle x\rangle_{y,N}$. Its mean is $\langle x\rangle = \mathbb{E}\left[\langle x\rangle_y\right]$ since

$$\boxed{\mathbb{E}\left[\langle x\rangle_{y,N}\right] = \mathbb{E}\left[\frac{1}{N}\sum_i \langle x_i\rangle_y\right] = \frac{1}{N}\sum_i \mathbb{E}\left[\langle x\rangle_y\right] = \mathbb{E}\left[\langle x\rangle_y\right]}$$

The variance of the sampling distribution is computed using the formula $\mathbb{E}\left[\langle x\rangle_{y,N}^2\right] - \mathbb{E}\left[\langle x\rangle_{y,N}\right]^2$. Then $\mathbb{E}\left[\langle x\rangle_{y,N}^2\right]$ is

$$\begin{aligned}
\mathbb{E}\left[\langle x\rangle_{y,N}^2\right] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_i \langle x_i\rangle_y\right)^2\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\sum_i\sum_j \langle x_i\rangle_y \langle x_j\rangle_y\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\sum_i \langle x_i\rangle_y^2 + \sum_i\sum_{j\neq i} \langle x_i\rangle_y \langle x_j\rangle_y\right] \\
&= \frac{1}{N^2}\sum_i \mathbb{E}\left[\langle x\rangle_y^2\right] + \frac{1}{N^2}\sum_i\sum_{j\neq i} \mathbb{E}\left[\langle x\rangle_y\right]^2 \\
&= \frac{1}{N}\mathbb{E}\left[\langle x\rangle_y^2\right] + \frac{N(N-1)}{N^2}\mathbb{E}\left[\langle x\rangle_y\right]^2.
\end{aligned}$$

Subtracting $\mathbb{E}\left[\langle x\rangle_{y,N}^2\right]$ from $\mathbb{E}\left[\langle x\rangle_{y,N}\right]^2$ gives the standard result

$$\boxed{\mathrm{Var}\left(\langle x\rangle_{y,N}\right) = \frac{1}{N}\mathbb{E}\left[\langle x\rangle_y^2\right] + \left(\frac{N(N-1)}{N^2} - 1\right)\mathbb{E}\left[\langle x\rangle_y\right]^2 = \frac{1}{N}\left(\mathbb{E}\left[\langle x\rangle_y^2\right] - \mathbb{E}\left[\langle x\rangle_y\right]^2\right).}$$

Sampling distribution for deviations from $\mathbb{E}\left[\langle\cdot\rangle_y\right]$

Approximating the variance of the sampling distribution for second-order central moments in terms of $\mathbb{E}\left[\langle\cdot\rangle_y\right]$ is slightly more complicated but follows the same procedure as the previous discussion. Without loss of generality the sampling distribution for $\langle u'c'\rangle_{y,N}$ will be derived

herein. First, its mean is

$$\begin{aligned}
\mathbb{E}\left[\langle u'c'\rangle_{y,N}\right] &= \mathbb{E}\left[\frac{1}{N}\sum_i \left\langle \left(u_i - \frac{1}{N}\sum_j \langle u_j\rangle_y\right)\left(c_i - \frac{1}{N}\sum_j \langle c_j\rangle_y\right)\right\rangle_y\right] \\
&= \frac{1}{N}\mathbb{E}\left[\left\langle \sum_i u_i c_i - \frac{1}{N}\sum_i\sum_j u_i \langle c_j\rangle_y - \dots \right.\right. \\
&\qquad \left.\left. \dots - \frac{1}{N}\sum_i\sum_j c_i \langle u_j\rangle_y + \frac{1}{N^2}\sum_i\sum_j \langle u_i\rangle_y \langle c_j\rangle_y \right\rangle_y\right] \\
&= \frac{1}{N}\sum_i \mathbb{E}\left[\langle uc\rangle_y\right] - \frac{1}{N^2}\mathbb{E}\left[\sum_i\sum_j \langle u_i\rangle_y \langle c_j\rangle_y\right] \\
&= \mathbb{E}\left[\langle uc\rangle_y\right] - \frac{1}{N^2}\mathbb{E}\left[\sum_i \langle u_i\rangle_y \langle c_i\rangle_y + \sum_i\sum_{j\neq i} \langle u_i\rangle_y \langle c_j\rangle_y\right] \\
&= \mathbb{E}\left[\langle uc\rangle_y\right] - \frac{1}{N^2}\sum_i \mathbb{E}\left[\langle u\rangle_y \langle c\rangle_y\right] - \frac{1}{N^2}\sum_i\sum_{j\neq i} \mathbb{E}\left[\langle u\rangle_y\right]\mathbb{E}\left[\langle c\rangle_y\right] \\
&= \mathbb{E}\left[\langle uc\rangle_y\right] - \frac{1}{N}\mathbb{E}\left[\langle u\rangle_y \langle c\rangle_y\right] - \frac{N(N-1)}{N^2}\mathbb{E}\left[\langle u\rangle_y\right]\mathbb{E}\left[\langle c\rangle_y\right]
\end{aligned}$$

Thus, to leading order, the mean for the sampling distribution of the dispersive flux is

$$\boxed{\mathbb{E}\left[\langle u'c'\rangle_{y,N}\right] = \mathbb{E}\left[\langle uc\rangle_y\right] - \mathbb{E}\left[\langle u\rangle_y\right]\mathbb{E}\left[\langle c\rangle_y\right] + \mathcal{O}(N^{-1}).} \tag{D.1}$$

To compute the variance of the sampling distribution, it is assumed without loss of

generality that $\mathbb{E}[u] = \mathbb{E}[c] = 0$ (and thus $\mathbb{E}\left[\langle u\rangle_y\right] = \mathbb{E}\left[\langle c\rangle_y\right] = 0$), since for $z \equiv x - \mathbb{E}[x]$

$$\begin{aligned}
\mathrm{Var}\left(\langle z\rangle_y\right) &= \mathbb{E}\left[\langle z\rangle_y^2\right] - \mathbb{E}\left[\langle z\rangle\right]^2 \\
&= \mathbb{E}\left[(\langle x\rangle_y - \mathbb{E}\left[\langle x\rangle_y\right])^2\right] \\
&= \mathbb{E}\left[\langle x\rangle_y^2 - 2\langle x\rangle_y \mathbb{E}\left[\langle x\rangle_y\right] + \left\langle \langle x\rangle_y \right\rangle^2\right] \\
&= \mathbb{E}\left[\langle x\rangle_y^2\right] - \mathbb{E}\left[\langle x\rangle_y\right]^2 \\
&= \mathrm{Var}\left(\langle x\rangle_y\right).
\end{aligned}$$

Applying the zero-mean assumption to the expression for the mean, (D.1), and keeping higher-order terms in N yields

$$\begin{aligned}
\mathbb{E}\left[\langle u'c'\rangle_{y,N}\right]^2 &= \left(\mathbb{E}\left[\langle uc\rangle_y\right] - \frac{1}{N}\mathbb{E}\left[\langle u\rangle_y \langle c\rangle_y\right]\right)^2 \\
&= \mathbb{E}\left[\langle uc\rangle_y\right]^2 - \frac{2}{N}\mathbb{E}\left[\langle uc\rangle_y\right]\mathbb{E}\left[\langle u\rangle_y \langle c\rangle_y\right] + \frac{1}{N^2}\mathbb{E}\left[\langle u\rangle_y \langle c\rangle_y\right]^2
\end{aligned}$$

Now,

$$
\begin{aligned}
\mathbb{E}\left[\langle u'c'\rangle_{y,N}^2\right] &= \mathbb{E}\left[\frac{1}{N^2}\left(\sum_i \left\langle \left(u_i - \frac{1}{N}\sum_j \langle u_j\rangle_y\right)\left(c_i - \frac{1}{N}\sum_j \langle c_j\rangle_y\right)\right\rangle_y\right)^2\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_i \left\langle u_i c_i - \frac{1}{N}u_i\sum_j \langle c_j\rangle_y - \frac{1}{N}c_i\sum_j \langle u_j\rangle_y + \dots \right.\right.\right. \\
&\qquad \left.\left.\left. \dots + \frac{1}{N^2}\sum_j\sum_k \langle u_j\rangle_y \langle c_k\rangle_y \right\rangle_y\right)^2\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_i \langle u_i c_i\rangle_y - \frac{1}{N}\sum_i \langle u_i\rangle_y \sum_j \langle c_j\rangle_y\right)^2\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\sum_{ij}\langle u_i c_i\rangle_y \langle u_j c_j\rangle_y - \frac{2}{N}\sum_{ijk}\langle u_i c_i\rangle_y \langle u_j\rangle_y \langle c_k\rangle_y + \dots \right. \\
&\qquad \left. \dots + \frac{1}{N^2}\sum_{ijkl}\langle u_i\rangle_y \langle c_j\rangle_y \langle u_k\rangle_y \langle c_l\rangle_y\right] \\
&= \underbrace{\frac{1}{N^2}\mathbb{E}\left[\sum_{ij}\langle u_i c_i\rangle_y \langle u_j c_j\rangle_y\right]}_{1} - \underbrace{\frac{2}{N^3}\mathbb{E}\left[\sum_{ijk}\langle u_i c_i\rangle_y \langle u_j\rangle_y \langle c_k\rangle_y\right]}_{2} + \dots \\
&\qquad \dots + \underbrace{\frac{1}{N^4}\mathbb{E}\left[\sum_{ijkl}\langle u_i\rangle_y \langle c_j\rangle_y \langle u_k\rangle_y \langle c_l\rangle_y\right]}_{3}
\end{aligned}
\tag{D.2}
$$

These terms will be expanded individually for simplicity. First,

$$
\begin{aligned}
\mathbb{E}\left[\sum_{ij}\langle u_i c_i\rangle_y \langle u_j c_j\rangle_y\right] &= \mathbb{E}\left[\overbrace{\sum_i \langle u_i c_i\rangle_y^2}^{(i=j)} + \overbrace{\sum_i\sum_{j\neq i}\langle u_i c_i\rangle_y \langle u_j c_j\rangle_y}^{(i\neq j)}\right] \\
&= \sum_i \mathbb{E}\left[\langle uc\rangle_y^2\right] + \sum_i\sum_{i\neq j}\mathbb{E}\left[\langle uc\rangle_y\right]^2.
\end{aligned}
$$

Note that the expectation is squared in the second term because the samples are independent, and the expectation of the product of independent variables is the product of the expecta-

tions. Furthermore, notice they are no longer indexed because the samples are identically distributed. Thus

$$\mathbb{E}\left[\sum_{ij}\left\langle u_i c_i\right\rangle_y\left\langle u_j c_j\right\rangle_y\right] = N\mathbb{E}\left[\left\langle uc\right\rangle_y^2\right] + N(N-1)\mathbb{E}\left[\left\langle uc\right\rangle_y\right]^2.$$

Similarly,

$$\begin{aligned}
\mathbb{E}\left[\sum_{ijk}\left\langle u_i c_i\right\rangle_y\left\langle u_j\right\rangle_y\left\langle c_k\right\rangle_y\right] = \mathbb{E}\Bigg[&\overbrace{\sum_i\left\langle u_i c_i\right\rangle_y\left\langle u_i\right\rangle_y\left\langle c_i\right\rangle_y}^{(i=j=k)} + \overbrace{\sum_i\sum_{j\neq i}\left\langle u_i c_i\right\rangle_y\left\langle u_j\right\rangle_y\left\langle c_j\right\rangle_y}^{(i\neq j=k)} + \dots \\
&\dots + \overbrace{\sum_i\sum_{k\neq i}\left\langle u_i c_i\right\rangle_y\left\langle u_i\right\rangle_y\left\langle c_k\right\rangle_y}^{(i=j\neq k)} + \overbrace{\sum_i\sum_{j\neq i}\left\langle u_i c_i\right\rangle_y\left\langle c_i\right\rangle_y\left\langle u_j\right\rangle_y}^{(i=k\neq j)}\Bigg] \\
&= N\mathbb{E}\left[\left\langle uc\right\rangle_y\left\langle u\right\rangle_y\left\langle c\right\rangle_y\right] + N(N-1)\mathbb{E}\left[\left\langle uc\right\rangle_y\right]\mathbb{E}\left[\left\langle u\right\rangle_y\left\langle c\right\rangle_y\right] + \dots \\
&\dots + N(N-1)\mathbb{E}\left[\left\langle uc\right\rangle_y\left\langle u\right\rangle_y\right]\mathbb{E}\left[\left\langle c\right\rangle_y\right] + \dots \\
&\dots + N(N-1)\mathbb{E}\left[\left\langle uc\right\rangle_y\left\langle c\right\rangle_y\right]\mathbb{E}\left[\left\langle u\right\rangle_y\right].
\end{aligned}$$

The assumption that $\mathbb{E}\left[\left\langle c\right\rangle_y\right] = \mathbb{E}\left[\left\langle u\right\rangle_y\right] = 0$, means that any term with an isolated first moment is zero. Then

$$\mathbb{E}\left[\sum_{ijk}\left\langle u_i c_i\right\rangle_y\left\langle u_j\right\rangle_y\left\langle c_k\right\rangle_y\right] = N\mathbb{E}\left[\left\langle uc\right\rangle_y\left\langle u\right\rangle_y\left\langle c\right\rangle_y\right] + N(N-1)\mathbb{E}\left[\left\langle uc\right\rangle_y\right]\mathbb{E}\left[\left\langle u\right\rangle_y\left\langle c\right\rangle_y\right].$$

From the previous term, note that term with a single index not equal to the others are the ones that evaluate to zero. Taking this into account for the last term, only those terms

that do not evaluate to zero are expanded out:

$$\mathbb{E}\left[\sum_{ijkl}\langle u_i\rangle_y\langle c_j\rangle_y\langle u_k\rangle_y\langle u_l\rangle_y\right] = \mathbb{E}\left[\overbrace{\sum_i\langle u_i\rangle_y^2\langle c_i\rangle_y^2}^{(i=j=k=l)} + \overbrace{\sum_i\sum_{k\neq i}\langle u_i\rangle_y\langle c_i\rangle_y\langle u_k\rangle_y\langle c_k\rangle_y}^{(i=j\neq k=l)} + \dots \right.$$
$$\left. \dots + \overbrace{\sum_i\sum_{j\neq i}\langle u_i\rangle_y^2\langle c_j\rangle_y^2}^{i=k\neq j=l} + \begin{smallmatrix} i\neq j=k=l \\ i=j=k\neq l \\ i=j=l\neq k \end{smallmatrix} \dots \right]$$
$$= N\mathbb{E}\left[\langle u\rangle_y^2\langle c\rangle_y^2\right] + N(N-1)\mathbb{E}\left[\langle u\rangle_y\langle c\rangle_y\right]^2 + \dots$$
$$\dots + N(N-1)\mathbb{E}\left[\langle u\rangle_y^2\right]\mathbb{E}\left[\langle c\rangle_y^2\right]$$

Substituting these expressions back into (D.2) gives

$$\mathbb{E}\left[\langle u'c'\rangle_{y,N}^2\right] = \frac{1}{N^2}\left(N\mathbb{E}\left[\langle uc\rangle_y^2\right] + N(N-1)\mathbb{E}\left[\langle uc\rangle_y\right]^2\right)\dots$$
$$-\frac{2}{N^3}\left(N\mathbb{E}\left[\langle uc\rangle_y\langle u\rangle_y\langle c\rangle_y\right] + N(N-1)\mathbb{E}\left[\langle uc\rangle_y\right]\mathbb{E}\left[\langle u\rangle_y\langle c\rangle_y\right]\right)\dots$$
$$+\frac{1}{N^4}\left(N\mathbb{E}\left[\langle u\rangle_y^2\langle c\rangle_y^2\right] + N(N-1)\mathbb{E}\left[\langle u\rangle_y\langle c\rangle_y\right]^2 + N(N-1)\mathbb{E}\left[\langle u\rangle_y^2\right]\mathbb{E}\left[\langle c\rangle_y^2\right]\right)$$

Isolating the terms that might cancel with $\mathbb{E}\left[\langle u'c'\rangle_{y,N}\right]^2$ ($\mathbb{E}\left[\langle uc\rangle_y\right]^2$, $\mathbb{E}\left[\langle uc\rangle_y\right]\mathbb{E}\left[\langle u\rangle_y\langle c\rangle_y\right]$, and $\mathbb{E}\left[\langle u\rangle_y\langle c\rangle_y\right]^2$) or are order at least N^{-1} gives

$$\mathbb{E}\left[\langle u'c'\rangle_{y,N}^2\right] = \frac{1}{N}\mathbb{E}\left[\langle uc\rangle_y^2\right] + \frac{N(N-1)}{N^2}\mathbb{E}\left[\langle uc\rangle_y\right]^2\dots$$
$$\dots - \frac{2N(N-1)}{N^3}\mathbb{E}\left[\langle uc\rangle_y\right]\mathbb{E}\left[\langle u\rangle_y\langle c\rangle_y\right] + \frac{N(N-1)}{N^4}\mathbb{E}\left[\langle u\rangle_y\langle c\rangle_y\right]^2 + o(N^{-1})$$

Finally, subtracting $\mathbb{E}\left[\langle uc\rangle_y\right]^2$ from $\mathbb{E}\left[\langle uc\rangle_y^2\right]$ yields

$$\begin{aligned}
\mathrm{Var}\left(\langle u'c'\rangle_{y,N}\right) &= \frac{1}{N}\mathbb{E}\left[\langle uc\rangle_y^2\right] + \left(\frac{N(N-1)}{N^2} - 1\right)\mathbb{E}\left[\langle uc\rangle_y\right]^2 \dots \\
&\dots - \left(\frac{2N(N-1)}{N^3} - \frac{2}{N}\right)\mathbb{E}\left[\langle uc\rangle_y\right]\mathbb{E}\left[\langle u\rangle_y\langle c\rangle_y\right] \dots \\
&\dots + \left(\frac{N(N-1)}{N^4} - \frac{1}{N^2}\right)\mathbb{E}\left[\langle u\rangle_y\langle c\rangle_y\right]^2 + o(N^{-1}) \\
&= \frac{1}{N}\mathbb{E}\left[\langle uc\rangle_y^2\right] - \frac{1}{N}\mathbb{E}\left[\langle uc\rangle_y\right]^2 + \frac{2}{N^2}\mathbb{E}\left[\langle uc\rangle_y\right]\mathbb{E}\left[\langle u\rangle_y\langle c\rangle_y\right] \dots \\
&\dots - \frac{1}{N^3}\mathbb{E}\left[\langle u\rangle_y\langle c\rangle_y\right]^2 + o(N^{-1})
\end{aligned}$$

$$\boxed{\mathrm{Var}\left(\langle u'c'\rangle_{y,N}\right) = \frac{1}{N}\left(\mathbb{E}\left[\langle uc\rangle_y^2\right] - \mathbb{E}\left[\langle uc\rangle_y\right]^2\right) + o(N^{-1}).}$$

Then the mean and variance of the sampling distribution of $\langle u'c'\rangle_{y,N}$ are

$$\boxed{\begin{aligned}
\mathbb{E}\left[\langle u'c'\rangle_{y,N}\right] &= \mathbb{E}\left[\langle uc\rangle_y\right] - \mathbb{E}\left[\langle u\rangle_y\right]\mathbb{E}\left[\langle c\rangle_y\right] + \mathcal{O}(N^{-1}), \\
\mathrm{Var}\left(\langle u'c'\rangle_{y,N}\right) &= \frac{1}{N}\left(\mathbb{E}\left[\langle uc\rangle_y^2\right] - \mathbb{E}\left[\langle uc\rangle_y\right]^2\right) + o(N^{-1}).
\end{aligned}} \tag{D.3}$$

Since centered moments of one variable are just the covariance of that variable with itself, the above expression simplifies for $\langle u'^2\rangle_{y,N}$ and $\langle c'^2\rangle_{y,N}$:

$$\boxed{\begin{aligned}
\mathbb{E}\left[\left\langle (u')^2\right\rangle_{y,N}\right] &= \mathbb{E}\left[\left\langle u^2\right\rangle_y\right] - \mathbb{E}\left[\langle u\rangle_y\right]^2 + \mathcal{O}(N^{-1}), \\
\mathrm{Var}\left(\left\langle (u')^2\right\rangle_{y,N}\right) &= \frac{1}{N}\left(\mathbb{E}\left[\left\langle u^2\right\rangle_y^2\right] - \mathbb{E}\left[\left\langle u^2\right\rangle_y\right]^2\right) + o(N^{-1}).
\end{aligned}} \tag{D.4}$$

D.2.2 Statistics of complex variables

At times the statistics of interest are in wavespace and are thus complex. Statistics for complex variables are similar to those for real variables, except that variances and covariances apply a complex conjugation to the second variable. For example, the covariance of two complex variables z_1 and z_2 is

$$\mathrm{Cov}\left(z_1, z_2\right) = \mathbb{E}\left[(z_1 - \langle z_1 \rangle)\overline{(z_2 - \langle z_z \rangle)}\right].$$

Mean and variance can be computed using the real and imaginary parts of the complex variable as

$$\mathbb{E}\left[z\right] = \mathbb{E}\left[\Re\left[z\right]\right] + i\mathbb{E}\left[\Im\left[z\right]\right],$$
$$\mathrm{Var}\left(z\right) = \mathrm{Var}\left(\Re\left[z\right]\right) + \mathrm{Var}\left(\Im\left[z\right]\right).$$

In cases where the real and imaginary parts of a complex variable are considered independently, the real and imaginary parts of the variance are also considered independently.

Appendix E

Scenario-dependence fits

This appendix details how scenario dependence was encoded in the distributions of the hyperparameters $\boldsymbol{\xi}$ of $p\big(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\xi}\big)$ by way of polynomial fits. The value of each hyperparameter was computed from the summary statistics of eigenvalue ensembles across a range of scenarios defined in terms of $\langle u'^2\rangle^{1/2}/\langle u\rangle$ and ℓ/L_x and was used generate the polynomial fits. For each hyperparameter a major direction of variation in $\langle u'^2\rangle^{1/2}/\langle u\rangle$ - ℓ/L_x space was identified. To make analysis more intuitive, the independent variables were rescaled so that their range was $\mathcal{O}(1)$. Since $\langle u'^2\rangle^{1/2}/\langle u\rangle$ ranged from around 0.5 to 1.5, no rescaling was necessary. Since ℓ/L_x ranged from around 0.02 to 1.0, it was rescaled by 0.08. Let the ordered pairs of independent variables be denoted $x_i = \Big(\langle u'^2\rangle^{1/2}/\langle u\rangle^{(i)}, (\ell/L_x(0.08)^{-1})^{(i)}\Big)$ and the corresponding computed hyperparameters be denoted y_i

Because the input space was only two-dimensional, it was possible to visually inspect surface and contour plots of each hyperparameter to identify the major direction of variation. See, e.g. Figure E.1. In higher dimensions the major direction could be found by performing a least-squares fit to a global linear model $\mathbf{y} = \mathbf{d}^T\mathbf{x} + c$ and projecting in the direction of $\mathbf{d}$, e.g. as described in [50]. Once the direction for each hyperparameter was identified, the computed values were projected in that direction. An example of the computed values projected in the major direction of variation as compared to another direction is shown in Figure E.2. By projecting in the major direction of variation, more collapse in the data is

apparent as compared to projecting onto a coordinate axis, for example.

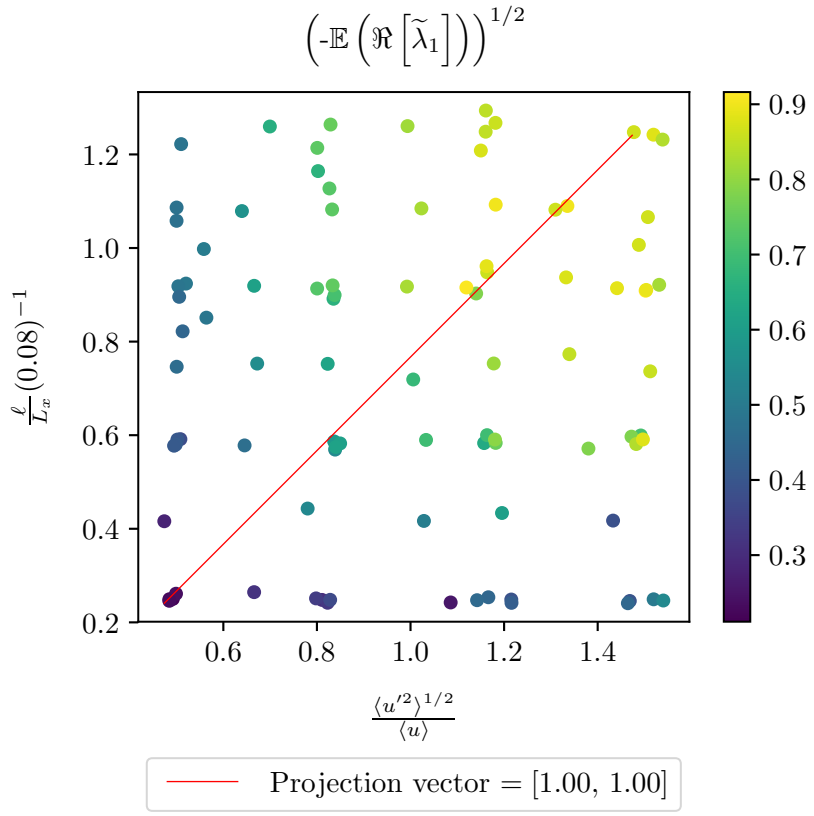

Figure E.1: Computed values for $\left(-\mathbb{E}\left(\Re\left[\widetilde{\lambda}_1\right]\right)\right)^{-1/2}$ across the scenario grid, with the most rapid direction of change.

Let the projected independent variable be denoted $s \equiv \mathbf{d}^T\mathbf{x}$. A one-dimensional polynomial was fit to the computed values as a function of s using a least-squares fit. For a given polynomial of order p, and N data points, the fitting problem is defined by the system of equations

$$y_i = c_0 + c_1 s_i + c_2 s_i^2 + \cdots + c_p s_i^p, i = 1, \cdots, N.$$

Converted into matrix form, the polynomial fit was computed by solving the least-squares

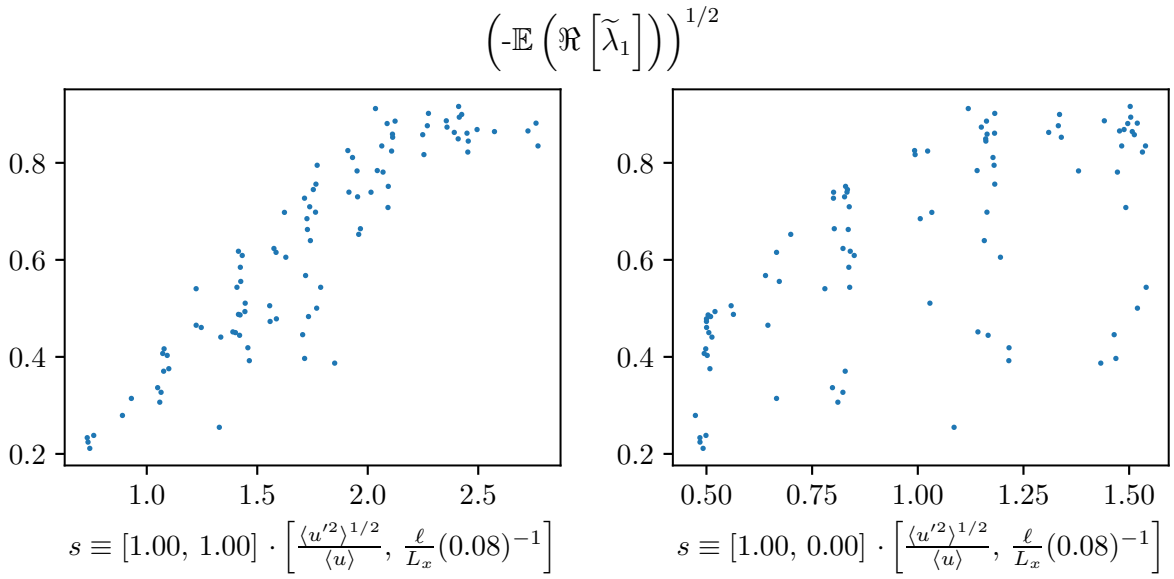

Figure E.2: Computed values of $\left(-\mathbb{E}\big(\Re[\widetilde{\lambda}_1]\big)\right)^{-1/2}$ projected in two different directions in scenario space, $[1, 1]$ and $[1.0]$.

minimization problem

$$\min_{\mathbf{c}} \left\| \mathbf{y} - \begin{bmatrix} | & | & | & | & | \\ \mathbf{1} & \mathbf{s} & \mathbf{s}^2 & \cdots & \mathbf{s}^p \\ | & | & | & | & | \end{bmatrix} \mathbf{c} \right\|_2^2 .$$

As discussed in Section 5.2, some hyperparameters are known to approach zero as $\langle u'^2\rangle^{1/2}/\langle u\rangle$ and ℓ/L_x approach zero. In that case the least-squares fit problem was instead defined as

$$\min_{\mathbf{c}} \left\| \mathbf{y} - \begin{bmatrix} | & | & | & | \\ \mathbf{s} & \mathbf{s}^2 & \cdots & \mathbf{s}^p \\ | & | & | & | \end{bmatrix} \mathbf{c} \right\|_2^2 ,$$

where the column of the least-squares matrix corresponding to the zeroth order has been removed.

This fit was used to define the mean for the hyperparameter distributions. The stan-

dard deviations of the distributions were computed based on the data's deviation from this modeled mean. The data points were not randomly sampled but were selected to provide an regularly-spaced covering of scenario space. Because of this, although data may appear to be less scattered at the extremes of the scenario space, this may just be due to the subset of scenarios that were used in the study. Care was taken that the scenario-dependent model of the standard deviation did not underestimate the uncertainty because of this.

For hyperparameters whose mean and variance were not required to approach zero along with $\langle u'^2 \rangle^{1/2} / \langle u \rangle$ and ℓ / L_x, the standard deviation was computed using the formula for a sample standard deviation using all the data and was assumed constant as a function of s. For the hyperparameters whose mean and variance must approach zero with $\langle u'^2 \rangle^{1/2} / \langle u \rangle$ and ℓ / L_x, the standard deviation was instead represented using a zero-intercept first-order polynomial fit. As shown in Figure E.3, raw deviations defined by $0.5\,|y_i - P(s_i)|$, where $P(s)$ is the polynomial fit, could vary widely for a given s.

The goal was to find a linear model that would encompass a majority of the deviations. First, to obtain data that would be more amenable to a least-squares fit, standard deviations were computed for ten overlapping bins spanning the range of s. The linear model was fit using the bin centers and binned standard deviations as data, up to and including the maximum standard deviation. As mentioned before, decreasing deviations at the edges of the scenario grid could be an artifact of approaching the edge of the scenario grid for which data was collected. Since in this case it is preferable to overestimate uncertainty, any bins at greater values of s than the maximum binned standard deviation were left out of the linear fit. An example of the binned standard deviations and the linear fit are shown in Figure E.3, and an example fit with scenario-dependent scenario deviation is shown in Figure E.4.

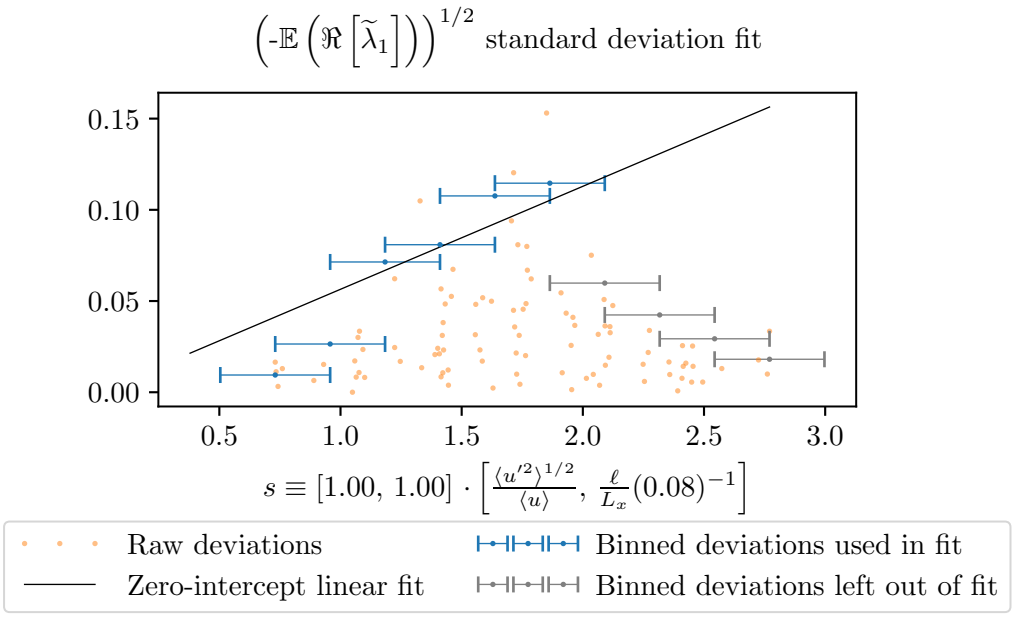

Figure E.3: A conservative estimate of the standard deviation with a linear model using only the binned standard deviations up to and including the maximum.

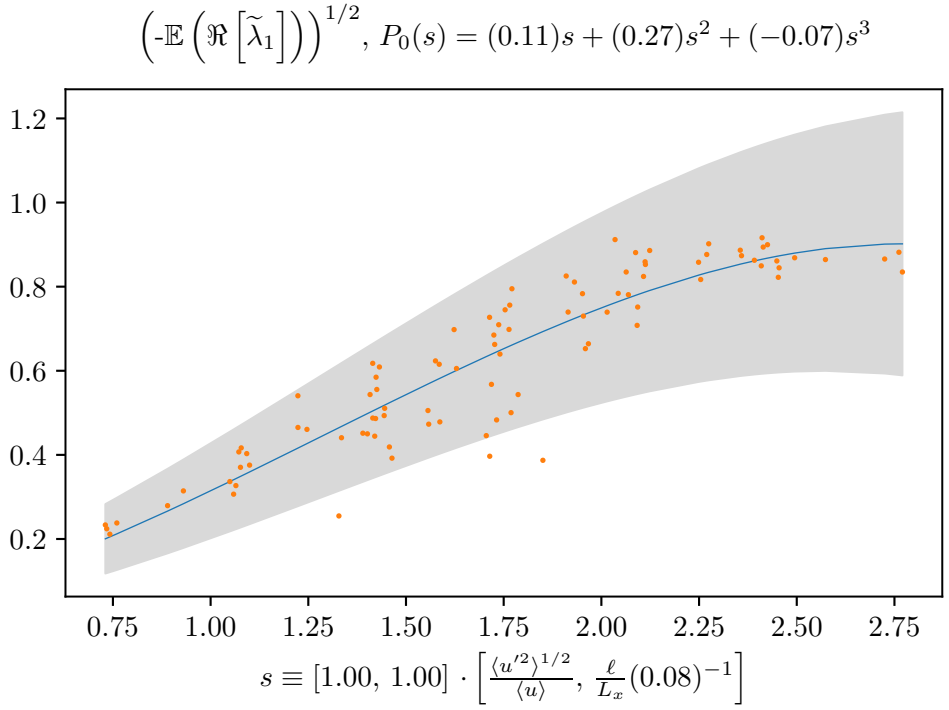

Figure E.4: A zero-intercept polynomial fit as a function of scenario for the mean and standard deviation of $\left(-\mathbb{E}\left(\Re\left[\widetilde{\lambda}_1\right]\right)\right)^{-1/2}$.

Appendix F

Full stochastic formulation of $\widetilde{\mathcal{L}}$

The stochastic formulation of $\widetilde{\mathcal{L}}$ is defined in terms of the distribution of its positive-wavenumber eigenvalues $\widetilde{\boldsymbol{\lambda}} = [\widetilde{\lambda}_1, \cdots, \widetilde{\lambda}_{N_k}]$ and hyperparameters $\boldsymbol{\xi}$, denoted $p(\widetilde{\boldsymbol{\lambda}}; \boldsymbol{\xi})$. N_k is the number of positive-wavenumber Fourier modes in the Fourier series expansion of the state variable $\langle c \rangle_y$, and $\widetilde{\lambda}_0 = 0$ to preserve mass. The full stochastic formulation of $p(\widetilde{\boldsymbol{\lambda}}; \boldsymbol{\xi})$ is

$$
\begin{aligned}
p(\widetilde{\boldsymbol{\lambda}}; \boldsymbol{\xi}) &= \mathcal{N}\Big(\mathbf{m}_{\boldsymbol{\lambda_R}} + i\mathbf{m}_{\boldsymbol{\lambda_I}}, \Sigma_2\Big), \\
\Sigma_2 &= w_1 \mathbf{v_1} \mathbf{v_1}^T + w_2 \mathbf{v_2} \mathbf{v_2}^T \\
\mathbf{v_1} &= [\mathbf{m}_{\mathbf{v_R}}\ \mathbf{m}_{\mathbf{v_I}}], \quad \mathbf{v_2} = [-\mathbf{m}_{\mathbf{v_I}}\ \mathbf{m}_{\mathbf{v_R}}], \\
(\mathbf{m}_\alpha)_k &= a_\alpha (k-1) + b_\alpha, \quad k \in [1, N_k], \quad \alpha \in \{\boldsymbol{\lambda_R}, \boldsymbol{\lambda_I}, \mathbf{v_R}, \mathbf{v_I}\},
\end{aligned}
$$

and $\boldsymbol{\xi} = \{a_{\boldsymbol{\lambda_R}}, b_{\boldsymbol{\lambda_R}}, a_{\boldsymbol{\lambda_I}}, b_{\boldsymbol{\lambda_I}}, w_1, w_2, a_{\mathbf{v_R}}, b_{\mathbf{v_R}}, a_{\mathbf{v_I}}, b_{\mathbf{v_I}}\}$. Let $P^{(i)}(s)$ denote a polynomial of degree i, and let a zero subscript denote a zero intercept. Let

$$
\begin{aligned}
s_{11} &= [1, 1] \cdot [\langle u'^2 \rangle^{1/2} / \langle u \rangle, \ell / L_x (0.08)^{-1}], \qquad s_{01} = [0, 1] \cdot [\langle u'^2 \rangle^{1/2} / \langle u \rangle, \ell / L_x (0.08)^{-1}], \\
s_{13} &= [1, 1/3] \cdot [\langle u'^2 \rangle^{1/2} / \langle u \rangle, \ell / L_x (0.08)^{-1}], \quad s_{10} = [1, 0] \cdot [\langle u'^2 \rangle^{1/2} / \langle u \rangle, \ell / L_x (0.08)^{-1}],
\end{aligned}
$$

The uncertain hyperparameters are distributed according to

$$
\begin{aligned}
(-a_{\boldsymbol{\lambda_R}})^{1/2} &\sim \mathcal{N}\left(P_0^3(s_{13}), P_0^1(s_{13})\right), & a_{\mathbf{v_R}} &\sim \mathcal{N}\left(P^2(s_{10}), P^0(s_{10})\right), \\
(-b_{\boldsymbol{\lambda_R}})^{1/2} &\sim \mathcal{N}\left(P_0^3(s_{11}), P_0^1(s_{11})\right), & b_{\mathbf{v_R}} &\sim \mathcal{N}\left(P^3(s_{11}), P^0(s_{11})\right), \\
a_{\boldsymbol{\lambda_I}} &\sim \mathcal{N}\left(P_0^3(s_{13}), P_0^1(s_{13})\right), & a_{\mathbf{v_I}} &\sim \mathcal{N}\left(P^2(s_{10}), P^0(s_{10})\right), \\
b_{\boldsymbol{\lambda_I}} &\sim \mathcal{N}\left(P_0^3(s_{11}), P_0^1(s_{11})\right), & b_{\mathbf{v_I}} &\sim \mathcal{N}\left(P^2(s_{11}), P^0(s_{11})\right), \\
(w_1)^{1/2} &\sim \mathcal{N}\left(P_0^2(s_{01}), P_0^1(s_{01})\right), & \left(\frac{w_1}{w_2}\right)^{1/2} &\sim \mathcal{N}\left(P^2(s_{11}), P^0(s_{11})\right),
\end{aligned}
$$

where the polynomial coefficients are determined according to the procedure described in Appendix E.

Bibliography

[1] Melissa Levy and Brian Berkowitz. Measurement and analysis of non-Fickian dispersion in heterogeneous porous media. *Journal of contaminant hydrology*, 64(3):203–226, 2003.

[2] Tian-Chyi Yeh, Raziuddin Khaleel, and Kenneth C Carroll. *Flow through heterogeneous geologic media*. Cambridge University Press, 2015.

[3] Shlomo P Neuman and Daniel M Tartakovsky. Perspective on theories of non-fickian transport in heterogeneous media. *Advances in Water Resources*, 32(5):670–680, 2009.

[4] Brian Berkowitz, Harvey Scher, and Stephen E Silliman. Anomalous transport in laboratory-scale, heterogeneous porous media. *Water Resources Research*, 36(1):149–158, 2000.

[5] SE Silliman and ES Simpson. Laboratory evidence of the scale effect in dispersion of solutes in porous media. *Water Resources Research*, 23(8):1667–1673, 1987.

[6] Lynn W Gelhar, Claire Welty, and Kenneth R Rehfeldt. A critical review of data on field-scale dispersion in aquifers. *Water resources research*, 28(7):1955–1974, 1992.

[7] Jacob Bear and Alexander H-D Cheng. *Modeling groundwater flow and contaminant transport*, volume 23. Springer Science & Business Media, 2010.

[8] Tian-Chyi Yeh, Raziuddin Khaleel, and Kenneth C Carroll. *Flow through heterogeneous geologic media*. Cambridge University Press, 2015.

[9] Shlomo P Neuman and Daniel M Tartakovsky. Perspective on theories of non-Fickian transport in heterogeneous media. *Advances in Water Resources*, 32(5):670–680, 2009.

[10] Todd A Oliver, Gabriel Terejanu, Christopher S Simmons, and Robert D Moser. Validating predictions of unobserved quantities. *Computer Methods in Applied Mechanics and Engineering*, 283:1310–1335, 2015.

[11] Marc C Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[12] K Sargsyan, HN Najm, and R Ghanem. On the statistical calibration of physical models. *International Journal of Chemical Kinetics*, 47(4):246–276, 2015.

[13] Rebecca E Morrison, Todd A Oliver, and Robert D Moser. Representing model inadequacy: A stochastic operator approach. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):457–496, 2018.

[14] Alberto Cialdea and Vladimir Maz'ya. *Semi-bounded differential operators, contractive semigroups and beyond*, volume 243. Springer, 2014.

[15] Bogdan Dumitrescu. *Positive trigonometric polynomials and signal processing applications*, volume 103. Springer, 2007.

[16] Tae Roh and Lieven Vandenberghe. Discrete transforms, semidefinite programming, and sum-of-squares representations of nonnegative polynomials. *SIAM Journal on Optimization*, 16(4):939–964, 2006.

[17] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Understanding and evaluating blind deconvolution algorithms. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1964–1971. IEEE, 2009.

[18] Michael J Daniels and Robert E Kass. Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, 94(448):1254–1263, 1999.

[19] Alan Huang, Matthew P Wand, et al. Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, 8(2):439–452, 2013.

[20] Benjamin Peherstorfer and Karen Willcox. Data-driven operator inference for nonintrusive projection-based model reduction. *Computer Methods in Applied Mechanics and Engineering*, 306:196–215, 2016.

[21] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

[22] Brian Berkowitz, Andrea Cortis, Marco Dentz, and Harvey Scher. Modeling non-Fickian transport in geological formations as a continuous time random walk. *Reviews of Geophysics*, 44(2), 2006.

[23] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[24] Ilya M Sobol. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280, 2001.

[25] Andrea Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer physics communications*, 145(2):280–297, 2002.

[26] Andrea Saltelli, Paola Annoni, Ivano Azzini, Francesca Campolongo, Marco Ratto, and Stefano Tarantola. Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Computer Physics Communications*, 181(2):259–270, 2010.

[27] Jonathan D Herman and Will Usher. SALib: An open-source Python library for Sensitivity Analysis. *J. Open Source Software*, 2(9):97, 2017.

[28] Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. Dram: efficient adaptive mcmc. *Statistics and computing*, 16(4):339–354, 2006.

[29] M Parno, P Conrad, A Davis, and YM Marzouk. MIT uncertainty quantification (MUQ) library, 2014.

[30] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[31] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[32] JP Delhomme. Spatial variability and uncertainty in groundwater flow parameters: a geostatistical approach. *Water Resources Research*, 15(2):269–280, 1979.

[33] Lynn W Gelhar and LW Gelhar. *Stochastic subsurface hydrology*, volume 390. Prentice-Hall Englewood Cliffs, NJ, 1993.

[34] Pol D Spanos and Roger Ghanem. Stochastic finite element expansion for random media. *Journal of engineering mechanics*, 115(5):1035–1053, 1989.

[35] Jun Zhang and Bruce Ellingwood. Orthogonal series expansions of random fields in reliability analysis. *Journal of Engineering Mechanics*, 120(12):2660–2677, 1994.

[36] A O'Hagan and Tom Leonard. Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*, 63(1):201–203, 1976.

[37] Robert D Moser. Kolmogorov inertial range spectra for inhomogeneous turbulence. *Physics of Fluids*, 6(2):794–801, 1994.

[38] Carl De Boor. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978.

[39] Richard W Johnson. Higher order B-spline collocation at the Greville abscissae. *Applied Numerical Mathematics*, 52(1):63–75, 2005.

[40] Mark Galassi, Jim Davies, James Theiler, Brian Gough, Gerard Jungman, Michael Booth, and Fabrice Rossi. GNU Scientific Library Reference Manual (Network Theory Ltd., 2009). *URL http://www. gnu. org/s/gsl*, 2009.

[41] Steven A Orszag. On the elimination of aliasing in finite-difference schemes by filtering high-wavenumber components. *Journal of the Atmospheric sciences*, 28(6):1074–1074, 1971.

[42] Matteo Frigo and Steven G Johnson. FFTW: An adaptive software architecture for the FFT. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 3, pages 1381–1384. IEEE, 1998.

[43] Myoungkyu Lee. *Direct numerical simulation (DNS) for incompressible turbulent channel flow at $Re_\tau = 5200$*. PhD thesis, The University of Texas at Austin, 2015.

[44] Philippe R Spalart, Robert D Moser, and Michael M Rogers. Spectral methods for the Navier-Stokes equations with one infinite and two periodic directions. *Journal of Computational Physics*, 96(2):297–324, 1991.

[45] Wai Y. Kwok. *Growth suppression in simulated compressible turbulent mixing layers*. PhD thesis, The University of Illinois at Urbana-Champaign, 2002.

[46] Stephen E. Guarini. *Direct numerical simulation of supersonic turbulent boundary layers*. PhD thesis, Stanford University, 1998.

[47] Rhys D. Ulerich. *Reducing Turbulence- and Transition-Driven Uncertainty in Aerothermodynamic Heating Predictions for Blunt-Bodied Reentry Vehicles*. PhD thesis, The University of Texas at Austin, 2014.

[48] Anders Logg, Kent-Andre Mardal, and Garth Wells. *Automated solution of differential equations by the finite element method: The FEniCS book*, volume 84. Springer Science & Business Media, 2012.

[49] Maurice G Kendall. *Advanced Theory Of Statistics Vol-I*. Charles Griffin: London, 6 edition, 1943.

[50] Paul G Constantine. *Active subspaces: Emerging ideas for dimension reduction in parameter studies*, volume 2. SIAM, 2015.

[51] Robert D Moser and Todd A Oliver. Validation of Physical Models in the Presence of Uncertainty. *Handbook of Uncertainty Quantification*, pages 129–156, 2017.

[52] Andrew M Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[53] Bamdad Hosseini and Nilima Nigam. Well-posed bayesian inverse problems: Priors with exponential tails. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):436–465, 2017.

[54] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

[55] Christopher K Wikle, Ralph F Milliff, Doug Nychka, and L Mark Berliner. Spatiotemporal hierarchical bayesian modeling tropical ocean surface winds. *Journal of the American Statistical Association*, 96(454):382–397, 2001.

[56] James Martin, Lucas C Wilcox, Carsten Burstedde, and Omar Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012.

[57] Heikki Haario, Eero Saksman, Johanna Tamminen, et al. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[58] Yves F Atchade. An adaptive version for the metropolis adjusted langevin algorithm with a truncated drift. *Methodology and Computing in applied Probability*, 8(2):235–254, 2006.

[59] Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[60] James Martin, Lucas C Wilcox, Carsten Burstedde, and Omar Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012.

[61] Noemi Petra, James Martin, Georg Stadler, and Omar Ghattas. A computational framework for infinite-dimensional Bayesian inverse problems, Part II: Stochastic Newton MCMC with application to ice sheet flow inverse problems. *SIAM Journal on Scientific Computing*, 36(4):A1525–A1555, 2014.

[62] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed August 2019].

[63] Gaël Guennebaud, Benoit Jacob, et al. Eigen v3. http://eigen.tuxfamily.org, 2010.

[64] The HDF Group. Hierarchical Data Format, v5. http://www.hdfgroup.org/HDF5/, 1997.

[65] J. D. Hunter. Matplotlib: A 2D graphics environment. 9(3):90–95, 2007.

[66] John H Cushman and Tim R Ginn. Fractional advection-dispersion equation: A classical mass balance with convolution-Fickian flux. *Water resources research*, 36(12):3763–3766, 2000.

[67] Rina Schumer, Mark M Meerschaert, and Boris Baeumer. Fractional advection-dispersion equations for modeling transport at the Earth surface. *Journal of Geophysical Research: Earth Surface*, 114(F4), 2009.

[68] Martin J Blunt, Branko Bijeljic, Hu Dong, Oussama Gharbi, Stefan Iglauer, Peyman Mostaghimi, Adriana Paluszny, and Christopher Pentland. Pore-scale imaging and modelling. *Advances in Water Resources*, 51:197–216, 2013.